THE UNIVERSITY *of* EDINBURGH

The evaluation of evidence for autocorrelated data, with an example relating to traces of cocaine on banknotes

Amy Wilson

Doctor of Philosophy
University of Edinburgh
2014

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Amy Wilson*)

Abstract

Much research in recent years for evidence evaluation in forensic science has focussed on methods for determining the likelihood ratio in various scenarios. One proposition concerning the evidence is put forward by the prosecution and another is put forward by the defence. The likelihood of each of these two propositions is calculated, given the evidence. The likelihood ratio, or value of the evidence, is then given by the ratio of the likelihoods associated with these two propositions.

The aim of this research is twofold. Firstly, it is intended to provide methodology for the evaluation of the likelihood ratio for continuous autocorrelated data. The likelihood ratio is evaluated for two such scenarios. The first is when the evidence consists of data which are autocorrelated at lag one. The second, an extension to this, is when the observed evidential data are also believed to be driven by an underlying latent Markov chain. Two models have been developed to take these attributes into account, an autoregressive model of order one and a hidden Markov model, which does not assume independence of adjacent data points conditional on the hidden states. A nonparametric model which does not make a parametric assumption about the data and which accounts for lag one autocorrelation is also developed. The performance of these three models is compared to the performance of a model which assumes independence of the data.

The second aim of the research is to develop models to evaluate evidence relating to traces of cocaine on banknotes, as measured by the log peak area of the ion count for cocaine product ion m/z 105, obtained using tandem mass spectrometry. Here, the prosecution proposition is that the banknotes are associated with a person who is involved with criminal activity relating to cocaine and the defence proposition is the converse, which is that the banknotes are associated with a person who is not involved with criminal activity relating to cocaine. Two data sets are available, one of banknotes seized in criminal investigations and associated with crime involving cocaine, and one of banknotes from general circulation. Previous methods for the evaluation of this evidence were concerned with the percentage of banknotes contaminated or assumed independence of measurements of quantities of cocaine on adjacent banknotes. It is known that nearly all banknotes have traces of cocaine on them and it was found that there was autocorrelation within samples of banknotes so these methods are not appropriate.

The models developed for autocorrelated data are applied to evidence relating to traces of cocaine on banknotes; the results obtained for each of the models are compared using rates of misleading

evidence, Tippett plots and scatter plots. It is found that the hidden Markov model is the best choice for the modelling of cocaine traces on banknotes because it has the lowest rate of misleading evidence and it also results in likelihood ratios which are large enough to give support to the prosecution proposition for some samples of banknotes seized from crime scenes. Comparison of the results obtained for models which take autocorrelation into account with the results obtained from the model which assumes independence indicate that not accounting for autocorrelation can result in the overstating of the likelihood ratio.

Lay summary

A forensic scientist in possession of a set of evidential data from a criminal case, such as measurements from a DNA sample, must assess the support given by these evidential data to a set of statements put forward as explanations for the evidence. Often there are two competing statements, known as propositions, one put forward by the prosecution and one put forward by the defence. The likelihood ratio has been developed as a method of providing a measure of the relative support given by the evidence to each of these two propositions. As an example, in the case where the evidence consists of measurements from a DNA sample, the prosecution proposition might be that the DNA belongs to the suspect and the defence proposition might be that the DNA belongs to some other unrelated person. The likelihood ratio for the two propositions, based on the evidence, can be used to obtain a number, representing the value of the evidence. If that number is greater than one, then the evidence is said to support the prosecution proposition. If the number is less than one, then the evidence is said to support the defence proposition. The logarithm of the likelihood ratio measures the extent of support for each of the two propositions. A large positive number gives strong support to the prosecution proposition. A small negative number (with large absolute value) gives strong support to the defence proposition.

Methods for the calculation of the likelihood ratio have been developed for many different types of evidential data. These methods are extended here for a new type of evidential data: data in which each measurement depends on the previous measurement taken. Examples of data of this sort are commonly found in finance. Consider a data set consisting of the value of inflation, measured in monthly intervals. The value of inflation in one month depends on the value of inflation in the previous month. If inflation is high in one month, then it is likely that inflation will still be high in the following month.

The motivating example for these new methods is evidence that relates to cocaine traces on banknotes. The amount of cocaine found on each of a sample of banknotes seized from a crime scene can be used in court as evidence of drug crime. The prosecution proposition is that the banknotes are associated with a person who is involved with drug crime relating to cocaine. The defence proposition is that the banknotes are associated with a person who is not involved with drug crime relating to cocaine. The problem with calculating the likelihood ratio for evidential data relating to traces of cocaine on banknotes is that it is known that cocaine can transfer from one banknote

to another. Therefore, the amount of cocaine found on one banknote in a sample depends on the amount of cocaine found on the banknotes either side of that banknote in the sample. A single highly contaminated banknote, introduced into a sample of notes by chance, might result in other banknotes in the sample becoming contaminated. Therefore, methods which allow each measurement to depend on the previous measurement must be used to calculate the likelihood ratio. Not accounting for this transfer of cocaine between banknotes might result in incorrect conclusions being drawn from analysis of the evidence, which could ultimately result in problems with the administration of justice.

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Colin Aitken, for his extensive support, guidance and dedication throughout this project. His encouragement and facilitation of my attendance at many conferences and meetings has also been hugely beneficial in providing ideas for this work. My second supervisor Dr Natalia Bochkina was also a source of useful ideas and encouragement. I am grateful to Dr Richard Sleeman of Mass Spec Analytical for many helpful discussions about the data and the propositions as well as for the provision of the data. For the generous funding of this work, I thank both Mass Spec Analytical and the EPSRC.

The current and former staff of Mass Spec Analytical have provided a great deal of support both during my three industrial placements, and in discussions throughout my PhD studies. In particular, Amy Johnson, Anna Milsom, Pete Luke, Deena Hooper and Beth Drake helped me with both access to and understanding of the data as well as giving me an understanding of the processes used for the testing of the banknotes, Jim Carter and Fay Thomas provided many interesting discussions about statistical problems in forensic analysis and Byron Morgan patiently taught me how to analyse banknotes using a mass spectrometer.

Many of the ideas for work in this thesis came from useful comments and discussions at meetings, workshops, conferences and visits. I am particularly grateful in this regard to the forensic statisticians at the Netherlands Forensic Institute, the University of Lausanne and the Institute of Forensic Research in Krakow, as well as Dr David Lucy at the University of Lancaster and Dr Tereza Neocleous at the University of Glasgow.

For helpful discussions on MCMC samplers I would like to thank Dr Richard Everitt and Dr Iain Murray. Anonymous referees for Forensic Science International and the Journal of the Royal Statistical Society (Applied Statistics) offered extremely helpful comments, particularly on the choice of propositions. For helpful advice on computing, programming, the production of software and for many useful discussions I would like to thank Adam Newgas, Hugo Vincent, Tim Gordon, Hari Sriskantha and Noel Hustler.

For a variety of support - statistical and otherwise - I would like to thank my friends, fellow PhD students, and family. Finally, I would like to thank Giles, for his support, encouragement and unwavering patience.

Contents

Chapter 1

Introduction

1.1 Introduction and motivation

Samples of banknotes can be seized from crime scenes as evidence of illegal activity involving cocaine. Cocaine traces on these banknotes can provide evidence to help determine whether or not the person with whom the banknotes are associated was involved with drug crime relating to cocaine. Methods have been developed to obtain an approximate measure of the amount of cocaine on each banknote within a sample of notes (Dixon et al. (2006); Sleeman et al. (2000)). Difficulties arise because it is known that the vast majority of banknotes from general circulation are also contaminated with cocaine (Jourdan et al. (2013); Jenkins (2001)), so scientific techniques which consider the quantity of cocaine found, rather than the number of contaminated banknotes, are required. It is known that heroin traces transfer between surfaces (Ebejer, Winn et al. (2007)) and it is suspected that the same is true for cocaine. Analysis of the available data confirmed this suspicion. If traces of cocaine can transfer between surfaces, this suggests that autocorrelation will be present between cocaine measurements within a sample of banknotes.

The aim of this thesis is twofold. Firstly, it is intended to develop models which effectively evaluate evidence relating to traces of cocaine on banknotes. The motivation behind this is to improve the administration of criminal justice in relation to this type of evidence. The evidence is evaluated using the likelihood ratio approach introduced in Lindley (1977), the benefits of which are discussed in Chapter 2. The exact form of the likelihood ratio used is not the same as that used in Lindley (1977), where likelihood ratios are used to compare the possible sources of a control and recovered item. This is because the problem being considered here involves only one evidential sample, a seized sample of banknotes. The likelihood ratio is used to provide a measure of support for whether this sample of banknotes is associated with a person who is, or is not, associated with drug crime relating to cocaine. More formally, the likelihood ratio is evaluated under two propositions. These are given by:

- H_C: the banknotes are associated with a person who is involved in crime involving cocaine.

- H_B: the banknotes are associated with a person who is not involved in crime involving cocaine.

The evidence is given by measurements of the quantity of cocaine on each banknote within a seized sample containing n banknotes. Denote this evidence by $\mathbf{z} = (z_1, \ldots, z_n)$. The likelihood ratio of the two propositions, H_C and H_B, for evidence $\mathbf{z}$ is given by

$$\text{LR} = \frac{f(\mathbf{z} \mid H_C)}{f(\mathbf{z} \mid H_B)}.$$

If this likelihood ratio is greater than one, then the evidence is said to support proposition H_C. If the likelihood ratio is less than one, then the evidence is said to support proposition H_B. This is as a result of the odds form of Bayes' theorem, given by

$$\frac{f(H_C \mid \mathbf{z})}{f(H_B \mid \mathbf{z})} = \text{LR} \times \frac{f(H_C)}{f(H_B)}.$$

The question that a jury, or other factfinder in a criminal case, ultimately wishes to answer is whether the accused has committed the crime with which he or she is charged. To evaluate the bearing of evidence from the seized banknotes on this question, the jury must discover which of the two propositions H_B or H_C is more likely, given the evidence $\mathbf{z}$. That is, given the cocaine traces found on the seized banknotes, is it more likely that the person with whom the banknotes are associated is involved with cocaine related crime than not. The answer to this question is determined by whether the left hand side of the odds form of Bayes' theorem (the posterior odds) is greater than or less than one. The posterior odds is given by the likelihood ratio multiplied by the prior odds, $f(H_C)/f(H_B)$. The likelihood ratio gives the factor for conversion of the prior odds to the posterior odds. If this factor is greater than one, then the likelihood ratio increases the support given by the prior odds for proposition H_C. If this factor is less than one, then the likelihood ratio decreases the support given by the prior odds for proposition H_C. In this way, the likelihood ratio can be used as a measure of support for one of the two propositions, H_C or H_B.

Four novel methods are used to evaluate the likelihood ratio. Three of these methods account for autocorrelation within samples of banknotes. Of these three methods, two are parametric approaches and one is a nonparametric approach. One parametric approach uses an autoregressive process of lag one to model the data. The other parametric approach uses a hidden Markov model. Banknotes are often stored in bundles, which may have come from a variety of different locations; it was found that banknotes in different bundles may have different levels of cocaine contamination. A hidden Markov model allows for the modelling of these different levels of contamination. The nonparametric approach dispenses with the Normality assumption required for the two parametric approaches, and models autocorrelation by estimating the conditional density function of the amount of cocaine measured on a banknote, conditional on the amount of cocaine measured on the previously analysed banknote in the same sample. An approach which does not account for autocorrelation is included for comparison. All of the approaches used are new approaches for the modelling of drug traces on

banknotes. In previous work in this area (discussed in Section 2.2), there has been no attempt to model autocorrelation, and no attempt to model the different levels of contamination found within samples of banknotes.

The four methods for the evaluation of the likelihood ratio for cocaine traces on banknotes are compared and conclusions are drawn about which model should be used in practice. Rates of misleading evidence are calculated to compare the methods. Evidence can be misleading in two ways. A sample of banknotes known to be associated with a person who is involved with crime involving cocaine could have a likelihood ratio smaller than one, so that the evidence suggests support for H_B. Alternatively, a sample of banknotes known to be associated with a person who is not involved with crime involving cocaine could have a likelihood ratio larger than one. These two different rates of misleading evidence can be used as one way of testing the models. Another way is to consider the actual values of the likelihood ratios obtained for samples of banknotes of known origin. To be useful a method must have low rates of misleading evidence, but it must also give likelihood ratios large enough so that support can be given to one of the propositions, where it is warranted. In addition, large likelihood ratios for samples of banknotes known to be associated with someone who is not involved with cocaine related crime are more strongly misleading than small likelihood ratios in this scenario.

The second aim for this thesis is to develop methodology for the evaluation of likelihood ratios for autocorrelated evidential data, of which the data relating to cocaine traces on banknotes are an example. Little work has been done on autocorrelated continuous evidential data (discussed in Section 2.1). There has been some work on autocorrelated discrete data (Aitken and Gold (2013)), and methods have been developed for evidential data which are multivariate Normal (so a full covariance matrix must be specified), but likelihood ratio methods for the purposes of evidence evaluation have not been applied to time series data which are autocorrelated. The three methods developed in this thesis to model the autocorrelated data relating to traces of cocaine on banknotes and evaluate the associated likelihood ratio, can also be used to evaluate the likelihood ratio for other autocorrelated evidential data, thus increasing the number of data types for which the likelihood ratio evidence interpretation framework can be applied. For example likelihood ratio techniques for autocorrelated continuous data will be required to evaluate evidence relating to traces of other types of illegal drug found on banknotes or for evidence consisting of drug traces found on other items, such as mobile phones or clothing.

1.2 Chapter summary

Chapter 2 reviews related literature in the field of forensic evidence evaluation. The chapter is split into two main sections, corresponding to the two aims of the thesis. The first section discusses methodology for evidence evaluation in the form of likelihood ratios, and summarises the data types for which these methods have been developed. The second section discusses evidence evaluation

relating specifically to traces of drugs on banknotes. The techniques in this second section are not all likelihood ratio techniques, and so the benefits and drawbacks of the various approaches used are summarised. At the end of Chapter 2, a summary of the novel developments introduced in this thesis is given.

Chapter 3 describes in detail three models for autocorrelated evidential data that are used to evaluate likelihood ratios in this thesis. The models are introduced generally, without discussion of data relating to traces of cocaine on banknotes. The three models are an autoregressive model with lag one (both with and without random effects), a hidden Markov model, and a nonparametric model, with two different bandwidth types. All three of these models can be used for evidential time series data where there is autocorrelation at lag one. The autoregressive and hidden Markov models make a Normality assumption, whereas the nonparametric model does not. The hidden Markov model is an extension to the autoregressive model, and allows for a more complicated data structure. It can be used for time series data which are autocorrelated, and where the parameters of the probability density function used to model each observation in the series (conditional on the previous observation) are determined by a latent Markov chain. Chapter 3 also describes how the two parametric models can be fitted to a training data set, using Bayesian techniques to obtain draws from the posterior distribution of the model parameters, conditional on the training data. Similarly, it is shown how nonparametric estimates of the probability density function of a sample of autocorrelated evidential data can be obtained from a training data set.

Chapter 4 contains methods for evaluating the likelihood ratio for the three models developed for general autocorrelated evidential time series data in Chapter 3. In addition, a method for evaluating the likelihood ratio for a standard model, which assumes independence between observations, is introduced. This is so that results both with and without the assumption of independence can be compared. Methods are described for the evaluation of the likelihood ratio when there is uncertainty over which of the two parametric models should be used. The between sample distribution, which gives the distribution of the model parameters across different time series, (this is discussed in further detail in Section 2.1) for the autoregressive model without random effects and the hidden Markov model is estimated using a weighted sum of the posterior distributions of the model parameters, obtained by conditioning on individual samples in training data sets. Similarly, the numerator and denominator of the likelihood ratio for the nonparametric model are estimated using a weighted sum of the estimated probability density functions of individual samples in training data sets, each evaluated for the new seized sample $\mathbf{z}$.

In Chapter 5, the data sets relating to traces of cocaine on banknotes are introduced. The data were collected during real forensic casework and so form a convenience sample. The techniques used to obtain measurements relating to traces of cocaine are described, and a novel peak detection algorithm, used to convert the raw data into a measure of the amount of cocaine on each banknote within a sample is introduced. In order to evaluate likelihood ratios for the data, two training data sets must be constructed. One, consisting of banknotes associated with proposition H_C, is used to

estimate the numerator of the likelihood ratio. The other, consisting of banknotes associated with proposition H_B, is used to estimate the denominator. The criteria used to select which samples belong to each of these two data sets are discussed, and alternatives considered. The formation of the data set associated with proposition H_C is novel; such a data set has not been used in previous work on evidence evaluation for traces of drugs on banknotes. There are, however, various difficulties associated with this data set. In particular, it was found that some samples in the data set associated with H_C are indistinguishable from samples in the data set associated with H_B. This difficulty, and the resulting effects on the evaluation of the evidence are discussed. Finally, the results following an exploratory data analysis of the samples in each of the two training data sets are presented, and the main attributes of the data are described.

Chapter 6 uses the methods introduced in Chapters 3 and 4 to evaluate the evidence relating to traces of cocaine on banknotes. Likelihood ratios are obtained for test data using each of the models described in Chapter 3. The models are then evaluated using rates of misleading evidence, Tippett plots and scatter plots, and conclusions are drawn as to the benefits and drawbacks of each of the models. Software, developed so that non-statisticians can easily and quickly apply the methods given in this thesis, is described.

Material in this thesis has been used in a modified form in Wilson et al. (2014) and Wilson et al. (2013).

Chapter 2

Literature review

The material in this chapter is divided into two main parts, corresponding to the two main aims of the research: to provide methodology for the evaluation of the likelihood ratio for autocorrelated evidential data, and to evaluate evidence relating to traces of cocaine on banknotes.

In the first part, the use of the likelihood ratio for evidence evaluation for different data types is discussed. The assumptions and limitations of the various approaches are presented. Two problems in forensic science are introduced: the comparison of sources problem, which compares whether a control and recovered item are likely to have originated from the same source, and the discrimination problem for determining from which of two populations a sample is more likely to have originated.

The second section reviews the literature relating to the use of drug traces on banknotes as evidence. Particular focus will be given to the statistical evaluation of such evidence. The limitations of the current approaches will be discussed, and the advantages of the use of the likelihood ratio approach for the evaluation of such evidence will be presented.

2.1 Evidence evaluation using likelihood ratios

2.1.1 The introduction of the likelihood ratio as a solution to the comparison of sources problem

Part of the role of a forensic scientist is to interpret evidence found at a crime scene, to aid factfinders in a criminal case (e.g. the judge or jury) in their decision making. The forensic scientist may be asked to comment on various competing statements about the evidence, each of which may be true or false. These statements, described here as propositions, can take many forms. They may refer to the immediate source of the evidence (e.g. the blood is from the suspect), an activity that has taken place (e.g. the suspect broke the window) or to an offence that has been committed (e.g. the suspect stole from the house). These different types of propositions are known respectively as source level, activity level and offence level propositions (Cook et al. (1998a); Evett, Jackson and Lambert

(2000)). Generally, a forensic scientist must consider two competing propositions relating to the evidence, one put forward by the prosecution in a criminal case, and one put forward by the defence (Cook et al. (1998b)). The aim of the forensic scientist is to evaluate the strength of support given by some evidential data (such as measurements from a DNA sample) to each of these two competing propositions.

A common problem occurs in forensic science when these propositions concern whether two objects are from the same source or from different sources. For example, if a glass fragment is found on a suspect and there is a broken window at the crime scene, one proposition might be that the glass fragment found on the suspect came from the window at the crime scene, and the other proposition might be that the glass fragment came from some other window. The evidence is given by a set of measurements from the glass fragment found on the suspect (known as the recovered sample) and a set of measurements from a glass fragment from the crime scene (known as the control sample).

Lindley (1977) developed a solution to this comparison of sources problem in the case where the measurements are univariate and are assumed to be independent and Normally distributed. Denote the n measurements on the control sample by $\mathbf{x} = (x_1, \ldots, x_n)$ and the m measurements on the recovered sample by $\mathbf{y} = (y_1, \ldots y_m)$. The corresponding means of each of these samples are denoted $\bar{x}$ and $\bar{y}$. The two propositions to be considered are

- H_p: the control and recovered sample are from the same source.
- H_d: the control and recovered sample are from different sources.

The proposition H_p has the subscript p because it is normally the prosecution proposition. Similarly the proposition H_d has the subscript d because it is normally the defence proposition. Lindley's solution assumes that the means $\bar{x}$ and $\bar{y}$ of the control and recovered sample follow Normal distributions with means θ_1 (control) and θ_2 (recovered) and variances σ^2/n (control) and σ^2/m (recovered). The two means θ_1 and θ_2 are also assumed to be Normally distributed, with mean μ and variance τ^2 (extensions are given to allow for a general non-Normal distribution, but the full analytical solution for the likelihood ratio cannot be derived in this case). Using a distribution for the means θ_1 and θ_2 in this way accounts for variance within source (σ^2) and variance between sources (τ^2). When proposition H_p is true, the means θ_1 and θ_2 are equal, because the control and recovered samples are from the same source.

The problem for the factfinder is to determine which of the two propositions (H_p or H_d) is more likely, given all of the evidence in the case. If the other evidence in the case is independent of the evidence $\mathbf{x}$ and $\mathbf{y}$ and is denoted by I, then this can be determined by considering the relative size of the two probabilities $P(H_p \mid \bar{x}, \bar{y}, I)$ and $P(H_d \mid \bar{x}, \bar{y}, I)$ (the means of the control and recovered sample are sufficient statistics so can be used in place of the measurements $\mathbf{x}$ and $\mathbf{y}$). The ratio of these two probabilities is known as the posterior odds. Let $f(\bar{x}, \bar{y} \mid H_p)$ be the joint probability density function of $\bar{x}$ and $\bar{y}$, given proposition H_p and let $f(\bar{x}, \bar{y} \mid H_d)$ be the joint probability density function of $\bar{x}$ and $\bar{y}$ given proposition H_d. By using the odds form of Bayes' theorem, the posterior odds can be written

$$\frac{P(H_p \mid \bar{x}, \bar{y}, I)}{P(H_d \mid \bar{x}, \bar{y}, I)} = \frac{f(\bar{x}, \bar{y} \mid H_p)}{f(\bar{x}, \bar{y} \mid H_d)} \frac{P(H_p \mid I)}{P(H_d \mid I)}.$$

The term

$$\frac{f(\bar{x}, \bar{y} \mid H_p)}{f(\bar{x}, \bar{y} \mid H_d)}$$

in this equation is known as the likelihood ratio, and the term

$$\frac{P(H_p \mid I)}{P(H_d \mid I)}$$

is known as the prior odds.

Lindley (1977) demonstrates that the likelihood ratio can be used to assess evidence in a criminal trial and hence is a solution to the comparison of sources problem. The likelihood ratio updates the prior odds, before consideration of the evidence $\bar{x}$ and $\bar{y}$, into the posterior odds, which take the new evidence into account. The posterior odds are the odds with which, ultimately, the factfinder is concerned. If the likelihood ratio multiplied by the prior odds is larger than one, then the probability of H_p given the evidence is larger than that of H_d given the evidence. It is the responsibility of the factfinder to determine a value for the prior odds, as they do not depend on the evidence $\bar{x}$ and $\bar{y}$. The posterior odds cannot be determined without the prior odds, so a forensic scientist, who is concerned only with the evidence $\bar{x}$ and $\bar{y}$ cannot usually comment on the value of the posterior odds. Instead, as discussed in Evett (1983), the forensic scientist's role is to provide the likelihood ratio, so that the prior odds can be updated with the information obtained from the new evidence. As such, the likelihood ratio can be considered as the strength of support of the evidence for one of the two propositions H_p or H_d.

Traditional hypothesis tests have also been used for evidence evaluation (an example relating to drugs on banknotes can be seen in Ebejer, Brereton et al. (2005)). However, the likelihood ratio approach has many advantages; a discussion of these can be seen in Aitken and Stoney (1991) and Aitken and Taroni (2004). One such advantage is that the likelihood ratio has no dependence on an arbitrary cut off point (e.g. 5% significance). Another advantage is that the use of a likelihood ratio ensures that a transposition of the conditional probabilities does not occur (known as the prosecutor's fallacy), a transposition which confuses the probability of finding the evidence on an innocent person with the probability of the innocence of a person on whom the evidence has been found. In addition, the likelihood ratio provides a method of comparing the likelihood of the evidence under the propositions of both the prosecution and the defence. This guards against potentially misleading situations when only one of these propositions is considered. Finally, an approach based on the likelihood ratio ensures equality of treatment of both propositions. In a procedure based on hypothesis testing, a null hypothesis is assumed true unless sufficient evidence is found to reject it at a pre-specified significance level.

Denoting the common mean of the measurements under the prosecution proposition by $\theta_1 = \theta_2 = \theta$, the likelihood ratio for the comparison of sources problem is given in Lindley (1977) by

$$\frac{\int f(\bar{x} \mid \theta) f(\bar{y} \mid \theta) f(\theta) \mathrm{d}\theta}{\int f(\bar{x} \mid \theta_1) f(\theta_1) \mathrm{d}\theta_1 \int f(\bar{y} \mid \theta_2) f(\theta_2) \mathrm{d}\theta_2}. \tag{2.1}$$

Lindley (1977) determines the analytical form of this likelihood ratio, given the independence and Normality assumptions detailed above. The density functions $f(\bar{x} \mid \theta)$ and $f(\bar{y} \mid \theta)$ are taken to be the density function of the Normal distribution. In later work (see, for example Aitken and Taroni (2004)) the distributions associated with these density functions are termed the within source distributions, because they account for the within source variance. The distribution associated with the density function $f(\theta)$ is termed the between source distribution, because it accounts for between source variance. The use of a between source distribution allows the rarity of the mean θ to be taken into account when assessing the strength of the evidence. If the control and recovered samples have similar means, and the mean is unusual, then the strength of evidence supporting the proposition that the samples are from the same source should be stronger than if the mean is relatively common.

2.1.2 The likelihood ratio approach for continuous evidential data

Lindley (1977) formulates the likelihood ratio for the comparison of sources problem for univariate measurements which are independent and Normally distributed. The analytical form of the likelihood ratio is derived when a Normal between source distribution is assumed. This between source distribution accounts for variation in the means of samples from different sources. Later work on evidence evaluation has extended the work done in Lindley (1977) to cover other data types, allowing for different forms of the within and between source distributions (Aitken and Lucy (2004); Aitken, Lucy et al. (2006); Aitken, Shen et al. (2007)). In Bozza et al. (2008) and Alberink et al. (2013), extensions are given so that the between source distribution in (2.1) becomes a function of both the mean and the variance. This allows for variation in the variance of samples from different sources. However, all of these extensions assume that the n measurements $\mathbf{x}$ are independent and that the m measurements $\mathbf{y}$ are independent; as a result, these methods cannot be used for autocorrelated data types, such as measurements associated with speech or drug traces on banknotes. This thesis addresses this shortfall by developing likelihood ratio methods for univariate continuous autocorrelated data.

In Aitken and Lucy (2004), the analytical form of the likelihood ratio is derived for multivariate measurements which are independent and which have a multivariate Normal distribution. The likelihood ratio is given for two forms of the between source distribution. The first assumes multivariate Normality, and the second uses nonparametric kernel density estimation to estimate the between source distribution. As in Lindley (1977) both of these methods only allow for the mean to vary between sources. Aitken, Zadora et al. (2007) use the kernel density approach given in Aitken and Lucy (2004) to calculate likelihood ratios for glass fragment data, with graphical models used to reduce the number of parameters needing to be estimated.

In Aitken, Lucy et al. (2006) the multivariate methods used in Aitken and Lucy (2004) are extended further to allow for another level of variance to be taken into account. The analytical form of the likelihood ratio for a three-level model is derived. Variation between the means of samples from different sources, variation between the means of different samples taken from the same source and variation within repeated measurements on the same sample are taken into account. This latter variation, not considered before, could arise as a result of measurement error. Assumptions of Normality and of there being no variation in the variance between sources are still made.

In Aitken, Shen et al. (2007) the assumption of Normality of the between source distribution of the mean is relaxed and an instead an exponential between source distribution is used. Like Lindley (1977), the between source distribution only allows for variation of the mean between samples, and the measurements on the control and recovered sample are univariate and assumed both independent and Normally distributed.

Assumptions that samples from different sources will have the same variance are relaxed in Bozza et al. (2008). As in Aitken and Lucy (2004), measurements are multivariate and independently Normally distributed but the between source distribution is taken to be the product of a multivariate Normal distribution (for the mean of the within source distribution) and an inverse Wishart distribution (for the covariance of the within source distribution). In this way, variation of covariances, as well as means, between different sources is taken into account. For this form of the between source distribution, it is not possible to obtain an analytical form of the likelihood ratio (Alberink et al. (2013)) so Markov chain Monte Carlo (MCMC) methods are used to estimate it. Bozza et al. (2008) writes the likelihood ratio for the comparison of sources problem in terms of marginal density functions as

$$\text{LR} = \frac{m(\mathbf{x}, \mathbf{y} \mid H_p)}{m(\mathbf{x} \mid H_d)\, m(\mathbf{y} \mid H_d)}$$

and uses the method described in Chib (1995) to estimate these marginal density functions. To do this, estimates of the probability density functions $f(\theta \mid \mathbf{x}, \mathbf{y}, H_p)$, $f(\theta \mid \mathbf{x}, H_d)$ and $f(\theta \mid \mathbf{y}, H_d)$ at the maximum likelihood estimate $\theta = \theta^*$ must be obtained. The parameter θ is multivariate because the between source distribution allows for variation in both the mean (denote by μ) and the covariance (denote by W). Bozza et al. (2008) assume that the estimate of the probability density function $f(\theta^* \mid \mathbf{x}, \mathbf{y}, H_p)$ can be factorised as

$$\hat{f}(\mu^*, W^* \mid \mathbf{x}, \mathbf{y}, H_p) = \hat{f}(\mu^* \mid \mathbf{x}, \mathbf{y}, H_p)\, \hat{f}(W^* \mid \mathbf{x}, \mathbf{y}, H_p)$$

where the parameter $\theta = (\mu, W)$. This assumes posterior independence of the parameters μ and W. Similar assumptions are made for the estimates $\hat{f}(\theta^* \mid \mathbf{x}, H_d)$ and $\hat{f}(\theta^* \mid \mathbf{y}, H_d)$. Alberink et al. (2013) notes that this assumption may be problematic when the independence assumption does not hold. However, this problem can be easily fixed, by noting that without an independence assumption

$$\hat{f}(\mu^*, W^* \mid \mathbf{x}, \mathbf{y}, H_p) = f(\mu^* \mid W^*, \mathbf{x}, \mathbf{y}, H_p)\, \hat{f}(W^* \mid \mathbf{x}, \mathbf{y}, H_p)$$

with the first term on the right hand side of the equation now conditional on the parameter W^*. The second term on the right can be estimated as in Bozza et al. (2008). The first term on the right is known analytically (and is given in Bozza et al. (2008)).

Alberink et al. (2013) use a similar approach to Bozza et al. (2008) to evaluate the likelihood ratio for the comparison of sources problem, in that variation in the variance parameter between sources is modelled as well as variation in the mean parameter, although in Alberink et al. (2013) the data are univariate. As with all of the other approaches discussed, the within source distribution is Normal, and the data are assumed independent. There are two main extensions seen in Alberink et al. (2013). The first is that three different distributions are used for the between source distribution. One is the univariate equivalent of the between source distribution used in Bozza et al. (2008) (called the semi-conjugate prior), one is a non-informative prior, proportional to the inverse of the variance, and one is the conjugate prior distribution seen on p. 74 of Gelman, Carlin et al. (2004). This conjugate prior distribution gives a between source distribution for the parameter $\theta = (\mu, \sigma^2)$ of

$$\mu \sim N(\mu_0, \sigma^2/\kappa_0)$$
$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

where μ_0, κ_0, ν_0 and σ_0^2 are hyperparameters to be estimated and the notation Inv-χ^2 corresponds to a scaled inverse chi-squared distribution. The difference between this and the univariate equivalent of the between source distribution used in Bozza et al. (2008) is that the variance of the parameter μ is proportional to σ^2. Alberink et al. (2013) derive the analytical form of the likelihood ratio for the two cases when the between source distribution is given by the non-informative prior and when the between source distribution is given by the conjugate prior.

As in Bozza et al. (2008), Alberink et al. (2013) use MCMC methods to evaluate the likelihood ratio when the between source distribution is given by the semi-conjugate prior, although there are differences in the implementation, leading to the second main extension. Alberink et al. (2013) use prior distributions on the hyperparameters of the between source distribution and then combine these prior distributions with training data to obtain a posterior distribution for the hyperparameters, conditional on the training data. All of the other methods discussed estimate the parameters of the between source distribution directly from the training data using summary statistics. The methods used in Alberink et al. (2013) allow for a Bayesian approach for the estimation of the between source distribution. One disadvantage of this approach is that the method for estimating the likelihood ratio used in Bozza et al. (2008) and introduced in Chib (1995) is no longer plausible, because instead of having a known analytic form for the between source density function, draws from the between source distribution are obtained using MCMC methods. Instead, Alberink et al. (2013) use Monte Carlo integration to estimate the likelihood ratio.

All of the literature discussed in this section evaluates likelihood ratios for continuous evidential

data. There are some common assumptions. All assume that measurements are independent and that the within source distribution is either the univariate or multivariate Normal distribution. Lindley (1977), Aitken and Lucy (2004), Aitken, Zadora et al. (2007) and Aitken, Lucy et al. (2006) allow for variation of the mean of the within source distribution between sources. Bozza et al. (2008) and Alberink et al. (2013) extend this by allowing the variance to vary between sources, which results in a need to use MCMC methods. Alberink et al. (2013) use a Bayesian approach to obtain the parameters of the between source distribution. There has so far been no consideration of the situation when measurements within the same source might be correlated. One way of modelling correlated univariate measurements would be to treat the n univariate control measurements $\mathbf{x} = (x_1, \ldots, x_n)$ and the m univariate recovered measurements $\mathbf{y} = (y_1, \ldots, y_m)$ as two multivariate measurements of dimension n and m respectively. Then the multivariate approaches used in Aitken and Lucy (2004) and Bozza et al. (2008) which use a multivariate Normal within source distribution could be used to model correlation between measurements through the covariance matrix. Adjustments, however, would be required because these two approaches assume that measurements have the same dimension. The difficulty with this technique would be the amount of training data required to estimate the covariance matrix. This thesis considers time series data which are autocorrelated. The correlation between the measurements within $\mathbf{x}$ and the measurements within $\mathbf{y}$ is modelled with a single autocorrelation parameter, which is allowed to vary between sources. The evaluation of likelihood ratios for this form of evidential data has not been considered before.

2.1.3 The likelihood ratio approach for autocorrelated discrete evidential data

The work discussed in the previous section evaluated likelihood ratios for continuous evidential data. Some work has also been done on the evaluation of evidence for discrete data, particularly in the field of DNA profiling (Buckleton et al. (2005)) and more recently on data relating to clicks in speech (Aitken and Gold (2013)). Both independent and autocorrelated discrete evidential data are considered in Aitken and Gold (2013), where the data are in the form of counts. The analytical form of the likelihood ratio for the comparison of sources problem is obtained in two scenarios. The first scenario assumes independent data and uses a Poisson within source distribution and a gamma between source distribution to model the variation of the mean of the observations between different sources. The second scenario concerns binary data and accounts for dependence between adjacent observations. The event that a count occurs in one time period is not assumed to be independent of the event that a count occurred in the previous time period. The likelihood ratio is derived for the situation where there are two time periods and two independent sets of observations of the two time periods (so that the control data are of the form $\mathbf{x} = \{x_{ij}; i \in \{1,2\}, j \in \{1,2\}, x_{ij} \in \{0,1\}\}$ for observation i and time period j with the recovered data defined similarly). The various conditional probabilities are specified and a beta distribution is used as the between source distribution of these conditional probabilities. The data considered in this thesis are continuous, so the models described in Aitken

and Gold (2013) cannot be used. However, Aitken and Gold (2013) do give a first step towards the consideration of autocorrelated evidential data, ideas which are expanded upon in this thesis.

2.1.4 Score-based likelihood ratio approaches

Score-based approaches for the evaluation of the likelihood ratio for the comparison of sources problem have been developed for situations where it is not clear how to formulate an exact probability model for the data. Score based approaches use a function to approximate the probability distribution function of a calculated 'score' which measures the similarity between the control and recovered sample. This score could be a function, such as a distance function, of the data from the control and recovered sample. An example would be the Euclidean distance between the control and recovered sample. Score based approaches have been used for fingerprint evidence (Neumann, Champod et al. (2007); Neumann, Evett et al. (2012)), handwriting (Davis et al. (2012)) and speech recognition (Gonzalez-Rodriguez et al. (2006)). Score-based methods do not require the distributional assumptions needed to fit the models in Section 2.1.2 (such as within source Normality) but do still require a function to be chosen to model the probability distribution function of the score. In addition, a suitable score must be chosen. In the context of autocorrelated data, if the control and recovered samples consist of autocorrelated measurements, and if the autocorrelation parameters are very discriminatory between samples from different sources, then careful consideration must be given to the choice of score. A score which ignores the autocorrelation in this case, such as the Euclidean distance between the first measurements in each of the control and recovered sample, might not be effective in providing likelihood ratios that accurately reflect the differences between control and recovered samples under each of the two propositions.

2.1.5 The likelihood ratio approach for the discrimination problem

The literature discussed so far gives different approaches for the evaluation of the likelihood ratio for the comparison of sources problem. This problem occurs when a control and recovered sample are obtained, and the question is whether these two samples are from the same source or not. This problem is one of the most common problems encountered in the forensic science literature on interpretation. The question considered in this thesis is different. The problem here concerns the determination of whether a sample of data is more likely from one population or another. Here, there is only one set of evidential data. A comparison between two objects is not being made. Instead, the aim is to decide which population the evidential data originate from. This problem will be termed a discrimination problem.

An example of the use of likelihood ratios in a problem of this sort can be seen in Evett, Pinchin et al. (1992), which considers whether a DNA sample is from someone of Afro-Caribbean or Caucasian ethnicity, and in Zadora et al. (2010), which looks at the discrimination of glass samples. As with the comparison of sources problem, the likelihood ratio alone cannot determine whether a set of data is

more likely from one population or another; it must be considered in conjunction with the prior odds. Chapter eight of Taroni, Bozza et al. (2010) discusses the derivation of the likelihood ratio for such discrimination problems. The likelihood ratio for a set of evidence consisting of n measurements, $\mathbf{z} = (z_1, \ldots, z_n)$, under two propositions, H_p and H_d, is considered. The two propositions are given by

- H_p : $\mathbf{z}$ are from population 1, and
- H_d : $\mathbf{z}$ are from population 2.

The likelihood ratio for the discrimination problem is given in Taroni, Bozza et al. (2010) by

$$\frac{f(\mathbf{z} \mid H_p)}{f(\mathbf{z} \mid H_d)}. \tag{2.2}$$

Let the parameter θ_1, possibly multivariate, characterise the probability density function of samples of data from population 1, and let the parameter θ_2 characterise the probability density function of samples of data from population 2. If the value of θ_i (for $i \in \{1,2\}$) varies between different samples in population i then by conditioning on θ_1 in the numerator and θ_2 in the denominator, the likelihood ratio can be written

$$\frac{\int f(\mathbf{z} \mid \theta_1) f(\theta_1) \mathrm{d}\theta_1}{\int f(\mathbf{z} \mid \theta_2) f(\theta_2) \mathrm{d}\theta_2}. \tag{2.3}$$

The probability density function $f(\theta_i)$ models the variability of the parameter θ_i between samples in population i, and hence will be termed the between sample density function (the associated distribution function will be termed the between sample distribution). This is analogous to the between source distribution used to model variability between sources in the comparison of sources problem. Similarly, the density function $f(\mathbf{z} \mid \theta_i)$, which is characterised by the parameter θ_i is termed the within sample density function (with the associated distribution function termed the within sample distribution).

Using this formulation for the likelihood ratio, the methods discussed previously for the evaluation of the likelihood ratio for the comparison of sources problem can be adapted to evaluate the value of evidence for discrimination problems. The limitations and assumptions of these methods will, however, still apply.

2.1.6 Conclusion

In this thesis, a discrimination problem is considered. The problem concerns the determination of which of two populations the continuous evidential data $\mathbf{z}$ originate. The likelihood ratio framework is used to calculate the value of the evidence $\mathbf{z}$, with reference to two propositions: that $\mathbf{z}$ came from population 1, and that $\mathbf{z}$ came from population 2. Most of the literature concerning the likelihood ratio framework for problems in forensic science focuses on the comparison of sources problem. However, the models used there can be extended to consider discrimination problems. For comparison of

sources problems, models have been developed for continuous data that assume a Normal within source distribution (either univariate or multivariate). Repeated measurements are assumed to be independent. This is often a sensible assumption but there are situations for which this assumption does not hold. For example, in data relating to speech, measurements taken in one time period may depend on measurements taken at other times. The data considered in this thesis relate to traces of cocaine on a sample of banknotes. The quantity of cocaine on one banknote may not be independent of the quantity of cocaine on adjacent banknotes. There is a need for methods to be developed for the evaluation of the likelihood ratio which dispense with the assumption of independence seen in the literature. This thesis addresses this gap.

2.2 Statistical methods for the evaluation of evidence relating to traces of drugs on banknotes

Little work has been done on building statistical models to evaluate evidence relating to drug traces on banknotes within the likelihood ratio framework. A notable exception is Besson (2004), with some of the methods used there also discussed and extended in Chapter eight of Taroni, Bozza et al. (2010). Several statistical methods outside of the likelihood ratio framework have been developed to evaluate this evidence. These include Lloyd (2009), Dixon et al. (2006), Ebejer, Brereton et al. (2005) and Jourdan et al. (2013). Methods developed in Lloyd (2009) and Dixon et al. (2006) are also discussed in Brereton (2009).

All of the above approaches involve making some statistical inference about the quantity of an illegal drug measured on a banknote or on each of a sample of banknotes seized from a crime scene. These measurements are the evidence, and are denoted by $\mathbf{z} = (z_1, \ldots, z_n)$ for a sample of n banknotes or just z for a single banknote. Besson (2004) and Dixon et al. (2006) approach the problem by providing a statistical method for the classification of $\mathbf{z}$ into one of two populations: banknotes from general circulation and banknotes from criminal cases. This is the discrimination problem described in Section 2.1.5, although a likelihood ratio approach is not taken in Dixon et al. (2006). Ebejer, Brereton et al. (2005), Jourdan et al. (2013) and Lloyd (2009) instead make inference by comparing $\mathbf{z}$ to just one population, that of general circulation banknotes. Banknotes are said either to be consistent with general circulation, or not. In the following two sections, the databases and statistical techniques used are summarised and compared.

2.2.1 Use of databases

All of the statistical approaches in the literature, likelihood ratio or otherwise, make use of a database of banknotes from general circulation. In addition, Besson (2004), Dixon et al. (2006) and Jourdan et al. (2013) consider a database of banknotes seized in the course of criminal investigations, although in Jourdan et al. (2013) it is not suggested that this should be used for inference. These seized 'crime'

banknotes are not necessarily associated with a conviction, they are just samples of banknotes seized by law enforcement agencies in the course of their casework. This leads to questions about the suitability of this 'crime' database for the discrimination of a new seized sample of banknotes in practice. If a new sample of banknotes is acquired which is found to be statistically similar (in some sense) to the database of 'crime' banknotes rather than the database of banknotes from general circulation, then what can be said about this new sample? It cannot be said that this new sample supports some proposition concerning involvement with crime because the database of 'crime' banknotes used to calculate this similarity are not known to be associated with crime, they have just been seized in the course of criminal investigation. In fact, because the new sample of banknotes are likely to have been seized themselves by law enforcement agencies in the course of criminal investigations, the new sample of banknotes are by definition part of the database of 'crime' banknotes. The calculation of any statistical similarity therefore adds nothing to the analysis because it is already known which of the two populations (general circulation or 'crime' banknotes) the newly seized banknotes belong to.

The difficulties associated with obtaining a database of banknotes that are truly associated with crime makes statistical approaches that are based on discrimination between two populations problematic. However, the consequences of not considering such a database can also be severe. Lloyd (2009), Ebejer, Brereton et al. (2005) and Jourdan et al. (2013) all base inference on a general circulation database only, so that if the drug contamination on a newly seized banknote (or newly seized sample of banknotes) is considered unusual in comparison to the contamination seen in the general circulation banknote database, this new banknote could be said to be 'inconsistent with the background population' (Ebejer, Brereton et al. (2005)). If a banknote is inconsistent with general circulation, it may not imply that this banknote is therefore associated with crime. As an extreme example, consider a banknote with no contamination of cocaine whatsoever (perhaps it is a new banknote). This measurement would be inconsistent with general circulation because the majority of banknotes in general circulation are contaminated with traces of cocaine (Jourdan et al. (2013); Jenkins (2001)), but it is incorrect to conclude that this banknote therefore must be associated with crime. As discussed on p. 101 of Aitken and Taroni (2004), for equal treatment of two propositions (that a banknote is from general circulation or that it is in some way associated with crime), it is necessary to consider two probabilities: the probability of obtaining the drug contamination found on the banknotes given that the banknotes are from general circulation and the probability of obtaining the drug contamination found on the banknotes given that the banknotes are associated with crime.

2.2.2 Statistical methods

The statistical techniques used by the literature in this section can be grouped into three different methodologies. Besson (2004) and Taroni, Aitken et al. (2006) use a likelihood ratio approach, Dixon et al. (2006) and Lloyd (2009) use a chemometric approach, reducing the dimension of the problem

with principal components analysis, and Ebejer, Brereton et al. (2005) and Jourdan et al. (2013) use an approach which involves modelling the probability density function of the drug contamination measured on banknotes from general circulation.

In Ebejer, Brereton et al. (2005), the drug in question is diamorphine (heroin), in the others, the drug discussed is cocaine. In Jourdan et al. (2013) and Besson (2004) the drug contamination on only one seized banknote is considered whereas Ebejer, Brereton et al. (2005), Lloyd (2009) and Dixon et al. (2006) consider the the drug contamination on a sample of banknotes. As discussed in Section 2.2.1, two databases are used to develop the methods. The database of banknotes from general circulation is denoted **x** and the database of banknotes seized by law enforcement agencies (where it is used) is denoted **y**.

In Dixon et al. (2006), principal components analysis (see Brereton (2009) for an introduction in this context) is used to reduce the dimension of the problem. The proportion of the n measurements **z** which fall into each of 51 equally spaced bins is used as a set of 51 variables. Similarly, a set of 51 variables for each of the samples in **x** and **y** is obtained. Principal components analysis is applied to reduce these 51 variables to two principal components. The method used is similar to that of principal components analysis for functional data, seen in Chapter eight of Ramsey and Silverman (2005); the functions being analysed are the probability density functions of the contamination on a banknote within each of the samples of banknotes in **x**, **y** and **z**. In Dixon et al. (2006), these probability density functions are discretised.

To classify the seized sample **z**, two Mahalanobis distances are calculated: one from the scores of the two principal components of **z** to the mean point of the scores of all samples in the database **x** and one from the scores of the two principal components of **z** to the mean point of the scores of all samples in **y**. The seized sample **z** is allocated to the database from which it has the smallest Mahalanobis distance. The two principal components used are selected by minimising the misclassification rate for seized samples of known origin.

In Lloyd (2009) the data are prepared as in Dixon et al. (2006), but using 29 equally spaced bins instead of 51. As before, principal components analysis is used on the data, and an optimal number of these principal components is selected. Then, three classification methods are compared: the Q-statistic, the D-statistic and support vector data description. These are described in detail in Lloyd (2009). They are one-class models, so they focus on whether a seized sample does, or does not, belong to the general circulation database **x**. No consideration is given to the database **y**. The three methods define statistics which can be used to form 95% probability intervals. Any seized sample with a statistic outside of the 95% probability interval is said to be different from general circulation. The method of preparing data relating to traces of cocaine on banknotes using principal components analysis, and the use of the methods in Lloyd (2009) and Dixon et al. (2006) with these data is also discussed in Brereton (2009).

The similarity in the approaches taken in Lloyd (2009) and Dixon et al. (2006) lies in the use of principal components analysis to transform the data. Use of principal components analysis reduces

the dimension of the problem being considered. Rather than considering the amount of cocaine on each banknote within a sample, this information is reduced to a small number of principal components. Reducing the dimension of the problem means not having to fit a statistical model to a vector of dimension n, as given by $\mathbf{z}$.

Jourdan et al. (2013) and Ebejer, Brereton et al. (2005) both estimate the probability density function of some statistic representing the drug contamination on samples of banknotes from general circulation. In Ebejer, Brereton et al. (2005), the statistic is a transformation of the proportion of contaminated banknotes within a sample. In Jourdan et al. (2013), the statistic is the amount of cocaine measured on a banknote. By using the proportion of contaminated banknotes within a sample, Ebejer, Brereton et al. (2005) avoid having to model the n-dimensional vector $\mathbf{z}$.

Ebejer, Brereton et al. (2005) consider the proportion of banknotes, within 48 samples from general circulation, that are contaminated with diamorphine. Two different transformations are applied to these proportions: an arcsin (square root) transformation and a log transformation. The Normal distribution is then used to model the transformed proportion. It is suggested that if the transformed proportion of contaminated banknotes in a seized sample is greater than the mean of the estimated Normal distribution plus three standard deviations then the seized sample may 'reasonably be considered as not consistent with those in general circulation and may, therefore, be associated with the handling or trafficking of heroin'.

Jourdan et al. (2013) take a similar approach. A power curve is fitted to the cocaine contamination found on banknotes from general circulation. This power curve is normalised to give a probability density function. On further inspection, the distribution fitted was found to be a Pareto distribution. Unlike in Dixon et al. (2006) and Lloyd (2009), logarithms are not taken of the data, so there are a large number of banknotes with very small quantities of contamination, meaning that a Normal distribution would not be appropriate. The probabilities of obtaining a banknote from general circulation with given ranges of contamination are calculated using this Pareto distribution. It is then said that the data 'allow the calculation of the probability that a bill is associated with illicit drug trafficking' and that this 'should prove useful in the evaluation of currency evidence in criminal cases'.

Both of these papers imply that if a seized sample is inconsistent with general circulation, then it therefore may be associated with drug crime. This is a flawed argument, sometimes termed the prosecutor's fallacy or the fallacy of the transposed conditional, because it arises from the confusion of conditional probabilities (see p79 and p112 of Aitken and Taroni (2004)). Let the statement that the banknotes are from general circulation be denoted by H_d, and the statement that the banknotes are associated with illicit drug trafficking by H_p. Let the evidence be denoted by E. In the case of Ebejer, Brereton et al. (2005), E is that the proportion of contaminated banknotes within a sample is greater than some given value. In the case of Jourdan et al. (2013), E is that the cocaine contamination z on a banknote falls within some given range. The actual probability considered in the two papers is $P(E \mid H_d)$, i.e. the probability of the evidence E occurring given that the banknotes are from general circulation. Ebejer, Brereton et al. (2005) imply that if this probability is small, then the sample may

'be associated with the handling or trafficking of heroin.'. This implies that $P(H_p \mid E)$ is large. However, the odds form of Bayes' theorem is given by

$$\frac{P(H_p \mid E)}{P(H_d \mid E)} = \frac{P(E \mid H_p)}{P(E \mid H_d)} \times \frac{P(H_p)}{P(H_d)}.$$

In order to infer information about $P(H_p \mid E)$ from $P(E \mid H_d)$, information about the prior probabilities $P(H_p)$ and $P(H_d)$, about the probability of the evidence assuming H_p, $P(E \mid H_p)$ and about the posterior probability $P(H_d \mid E)$ are needed. The conditional probability $P(E \mid H_d)$ has been confused with the conditional probability $P(H_d \mid E)$. If the statements H_d and H_p are assumed to be exhaustive then knowledge of $P(H_d \mid E)$ would imply knowledge of $P(H_p \mid E)$.

Similarly, Jourdan et al. (2013) implies that the probability $P(E \mid H_d)$ can give the probability that a 'bill is associated with illicit drug trafficking'. This latter probability is again given by $P(H_p \mid E)$, so again can not be inferred only from the probability $P(E \mid H_d)$.

In Besson (2004), these concerns over the interpretation of the probabilities calculated are resolved by use of the likelihood ratio approach, as discussed in detail in Section 2.1. The approach in Besson (2004) gives consideration to two propositions:

- H_p : the banknotes are involved in drug trafficking ,and
- H_d : the banknotes are from general circulation.

Use of two propositions avoids the problems seen in the approaches of Jourdan et al. (2013) and Ebejer, Brereton et al. (2005), where the first of these two propositions is ignored. The use of a likelihood ratio allows for a value to be obtained for the weight of evidence, which can be any real number. Larger numbers imply greater support for proposition H_p. This is an improvement on the methods used in Dixon et al. (2006) and Lloyd (2009), which have a binary result. Either the seized sample, $\mathbf{z}$ is classified as being from the general circulation database, or it is classified as being from the database of seized samples (or in the case of Lloyd (2009), it is classified as not being from the general circulation database). It can be argued that in evidence evaluation, having a scale of values is more useful than a binary classification as it allows for a measure of the extent to which a proposition is supported.

In Besson (2004) methods are presented to evaluate the likelihood ratio both for a single seized banknote and for the proportion of contaminated banknotes within a sample. The method for a single seized banknote divides the possible quantities of cocaine contamination on a banknote into discrete intervals. The likelihood ratio for a new seized banknote with contamination z, falling in one of these intervals, is calculated by dividing the proportion of banknotes from database $\mathbf{y}$ with contamination in that interval by the proportion of banknotes from database $\mathbf{x}$ with contamination in that interval. This method puts a discrete probability mass function on the intervals of contamination so that the likelihood ratio, analogous to the continuous version in (2.2), is given by

$$\text{LR} = \frac{P(z \mid H_p)}{P(z \mid H_d)}.$$

The probability $P(z \mid H_p)$ is estimated by the proportion of banknotes from the 'crime' database (**y**) with contamination in the same interval as the seized banknote and the probability $P(z \mid H_d)$ is estimated by the proportion of banknotes from the general circulation database (**x**) with contamination in the same interval as the seized banknote.

Another approach in Besson (2004) uses the proportion of contaminated banknotes in a seized sample as the evidence, as done for heroin contamination in Ebejer, Brereton et al. (2005). The proportion of contaminated banknotes within a sample is modelled using a binomial distribution, the parameters of which are estimated from the databases **x** (for proposition H_d) and **y** (for proposition H_p). This approach is expanded on in Taroni, Aitken et al. (2006). In Taroni, Aitken et al. (2006), The probability parameter of the binomial distribution is allowed to vary between samples; a beta distribution is used for this between sample distribution. As noted in Taroni, Aitken et al. (2006), this approach requires an assumption of independence between quantities of contamination on banknotes within the same sample. A problem with using this approach in practice is that almost all banknotes from general circulation are contaminated with cocaine (Jenkins (2001); Dixon et al. (2006); Jourdan et al. (2013)), so it is difficult to discriminate between samples of banknotes from general circulation and samples of banknotes associated with crime using only the proportion of contaminated banknotes within a sample. This problem is noted in Besson (2004).

In Taroni, Aitken et al. (2006), a method is presented for calculating the likelihood ratio where the evidence consists of measurements on a whole sample of banknotes instead of a single banknote. All of the other methods presented in this section have used either a single banknote as the evidence, the proportion of contaminated banknotes within a sample, or have transformed the full set of evidential data **z** and used a subset of the transformed data as the evidence. To evaluate the likelihood ratio for the n measurements **z**, kernel density estimates of both the probability density function of quantities of contamination on 'crime' banknotes (**y**) and the probability density function of quantities of contamination on general circulation banknotes (**x**) are calculated. The likelihood ratio for **z** is given in Taroni, Aitken et al. (2006) by

$$\text{LR} = \frac{\hat{f}(z_1 \mid H_p)\hat{f}(z_2 \mid H_p)\ldots\hat{f}(z_n \mid H_p)}{\hat{f}(z_1 \mid H_d)\hat{f}(z_2 \mid H_d)\ldots\hat{f}(z_n \mid H_d)} \tag{2.4}$$

where $\hat{f}(\cdot \mid H_p)$ is the kernel density estimate of the probability density function of quantities of contamination on 'crime' banknotes and $\hat{f}(\cdot \mid H_d)$ is the kernel density estimate of the probability density function of quantities of contamination on general circulation banknotes. Equation (2.4) is equivalent to the product of n likelihood ratios as given in (2.2), with one likelihood ratio for the measurements on each of the individual banknotes in **z**. In order to write the likelihood ratio as the product of the likelihood ratios for the individual banknotes, an assumption of independence

between quantities of contamination on banknotes within the seized sample, **z**, has been made. It is noted in Taroni, Aitken et al. (2006) that this assumption may be unreasonable in practice.

2.2.3 Conclusion

Jourdan et al. (2013) and Ebejer, Brereton et al. (2005) have developed an approach based on the estimation of the probability density function of banknotes from general circulation, for the evaluation of drug traces on banknotes. Difficulties arise with the interpretation of the conclusions drawn from these approaches because of the risk of committing the prosecutor's fallacy. The consideration of just one database of banknotes, those from general circulation, as done in Lloyd (2009) could result in similar interpretation problems. Lloyd (2009) and Dixon et al. (2006) use chemometric methods to evaluate the evidence, reducing the dimensionality of the problem by using principal components analysis to discard some of the information. The methods used have a binary conclusion, the seized sample **z** is determined to have come from one population or the other. Besson (2004) and Taroni, Aitken et al. (2006) use a likelihood ratio approach which resolves issues around the interpretation of the results, and also allows for a weight of evidence to be calculated. In addition, Taroni, Aitken et al. (2006) provide methods for the evaluation of the measurements of a full sample of banknotes, rather than for a single note (Jourdan et al. (2013); Besson (2004)), the proportion of contaminated notes within a sample (Besson (2004); Ebejer, Brereton et al. (2005)) or by discarding some of the information (Lloyd (2009); Dixon et al. (2006)). However, the approach of Taroni, Aitken et al. (2006) is limited by the assumption of independence between quantities of contamination on banknotes within the same sample.

2.3 The novelty of the present work

As discussed in Section 2.2.2, the likelihood ratio approaches of Besson (2004) and Taroni, Aitken et al. (2006) avoid the problems of interpretation seen in Ebejer, Brereton et al. (2005) and Jourdan et al. (2013). Nevertheless, there remain difficulties with the application of the approaches in Besson (2004) and Taroni, Aitken et al. (2006) to data relating to traces of cocaine on banknotes. In Besson (2004), the evidence for a single banknote is evaluated. In Taroni, Aitken et al. (2006), this is extended to a sample of banknotes, but only in the case where the measurements on a sample of banknotes are independent of one another. In Ebejer, Winn et al. (2007), it is shown that heroin transfers from one banknote to another. It is thought that cocaine also transfers between notes. Various studies have shown that over 90% of banknotes from general circulation are contaminated with cocaine (Jourdan et al. (2013); Jenkins (2001)). This would support the view that cocaine transfers between banknotes because it is unlikely that all of these general circulation banknotes have been contaminated as a result of their involvement in drug use or drug crime. It is more likely that banknotes in general circulation obtain their contamination from being in contact with other contaminated objects. It

might therefore be the case that an assumption of independence between the measurements on a sample of banknotes does not hold. In order to evaluate the evidence associated with a full sample of banknotes (as in Taroni, Aitken et al. (2006)) without making an assumption of independence, correlation between the measurements of cocaine associated with these banknotes must be taken into account.

The work done in this thesis builds on the work done in Besson (2004) and Taroni, Aitken et al. (2006) with three novel developments. The first of these novel developments is to dispense with the assumption of independence. An autocorrelation parameter is introduced to model autocorrelation between measurements on a sample of banknotes. As the measurements are not assumed to be independent, (2.4) cannot be used to evaluate the likelihood ratio, so new methods that account for autocorrelation are given. The likelihood ratio approaches described in Section 2.1.2 for the evaluation of evidence for continuous data all either assume independence of measurements, or assume that measurements are multivariate Normal, and so require the specification of a full covariance matrix. Therefore, the methods developed for data relating to traces of cocaine on banknotes, where the measurements are autocorrelated, are also novel for the evaluation of other types of evidential data.

The second development is to model cocaine traces on banknotes using a hidden Markov model. It was found when analysing the data that samples of banknotes (particularly those associated with someone who is involved with crime) often consisted of multiple bundles, each with a different pattern of contamination. Banknotes within the same bundle had similar levels of contamination. It was found that this effect, combined with the autocorrelation, could be modelled with a hidden Markov model. Within the literature relating to traces of drugs on banknotes, this pattern of contamination involving different bundles has not been found or modelled before. As with the inclusion of autocorrelation, the use of a hidden Markov model to obtain likelihood ratios for continuous evidential data is also new within the forensic statistics literature.

The third development relates to the choice of database. As noted in Section 2.2.1, there are problems with the database used to represent samples of banknotes associated with crime. In this thesis, a more appropriate database is compiled, which consists of samples of banknotes that were associated with a criminal case which eventually resulted in a conviction of a crime involving cocaine (either through a trial or through a guilty plea).

The likelihood ratio approaches of Besson (2004) and Taroni, Aitken et al. (2006) provide the most logically sound methodology of the approaches discussed in Section 2.2.2 for the evaluation of traces of cocaine on banknotes. This thesis extends the approaches of Besson (2004) and Taroni, Aitken et al. (2006) by addressing several of their shortfalls: by dispensing with the assumption of independence; by allowing for the modelling of different levels of contamination on different bundles of banknotes within one sample; and by the use of a more appropriate database. These developments will be discussed in the chapters which follow.

Chapter 3

Modelling autocorrelated evidential data

In this chapter, three different models that can be used for autocorrelated data are presented. The autocorrelated data considered are sequential, continuous, time series data, measured at equally spaced discrete time points. Two of the three models presented are parametric, and one is nonparametric. A description of each of these three models is given, along with a discussion of the fitting of these models to a data set.

The ultimate aim of this work is to use the fitted models presented in this chapter to obtain likelihood ratios for two different propositions concerned with the origin of a new (unseen) set of autocorrelated data, so that the weight of evidence of this new data can be evaluated. More specifically, given two sets of training data, B and C, each containing multiple samples of autocorrelated data, it is desired to obtain parameter estimates (in the case of a parametric model) or function estimates (in the case of a nonparametric model) for each of the models, using each of these sets, B and C, separately. These parameter or function estimates can then be used to evaluate the likelihood ratio for a new set of evidential data, say $\mathbf{z}$, for the two propositions: H_B, that the data $\mathbf{z}$ are from the set B and H_C, that the data $\mathbf{z}$ are from the set C. The subscripts B and C are chosen to represent 'background' and 'crime', because a standard set of propositions might be H_B, that the data $\mathbf{z}$ are from the background population, and H_C, that the data $\mathbf{z}$ are associated with crime. Other propositions could be used if there were corresponding sets of training data. In this chapter, the fitting of the models is described. Methods for obtaining likelihood ratios are discussed in Chapter 4.

The three different models described in this chapter are an autoregressive process of order one, a hidden Markov model and a nonparametric model which uses kernel density estimates of conditional density functions. Two different methods of bandwidth selection are used in the fitting of the kernel density estimates. The autoregressive process allows for autocorrelation between adjacent observations. Two different methods are described for the modelling of this autoregressive process:

one uses random effects and one uses fixed effects. The hidden Markov model is an extension to the autoregressive process which, in addition to accounting for autocorrelation, also allows for different parameter sets to be used for different observations. The choice of which parameter set to use is governed by a latent variable. Each observation is associated with one of these latent variables. The latent variables form a Markov chain, so that there is a dependence of the parameter set used by an observation on the parameter set used by the previous observation. The nonparametric model uses kernel density estimates to evaluate the density function for an observation, conditioned on the previous observation. These kernel density estimates are computed using both a fixed and a variable bandwidth.

Before describing the models some background and notation is introduced more formally. It is thought that a crime has been committed. Part of the evidence involves a sample of autocorrelated data. There are two competing propositions for the origin of the data (e.g. associated with crime or from the background population), with two training sets of data, one for each of the propositions. To evaluate the weight of evidence, the likelihood ratio associated with these two propositions must be calculated. The two training sets, used to develop the models for the calculation of the likelihood ratio, are

- The set B, containing continuous data $\mathbf{x} = \{x_{it}; i = 1,\ldots,m_B, t = 1,\ldots,n_{B_i}\}$: there are m_B independent samples with n_{B_i} autocorrelated observations in sample i. Note that B refers to data from the background population. The notation $\mathbf{x}_i$ is used to indicate the ith sample, $x_{i1},\ldots,x_{in_{B_i}}$.
- The set C, containing continuous data $\mathbf{y} = \{y_{it}; i = 1,\ldots,m_C, t = 1,\ldots,n_{C_i}\}$: there are m_C independent samples with n_{C_i} autocorrelated observations in sample i. Note that C refers to data associated with crime. The notation $\mathbf{y}_i$ is used to indicate the ith sample, $y_{i1},\ldots,y_{in_{C_i}}$.

The questioned sample is

- $\mathbf{z} = (z_1, z_2, \ldots, z_n)$: a sample of n autocorrelated observations of unknown and questioned origin. This questioned sample may sometimes be known as the seized sample as it has been seized by law enforcement agencies.

The two propositions concerning the origin of the sample of unknown origin (the questioned sample) are

- H_B: the questioned sample is from the set B.
- H_C: the questioned sample is from the set C.

Often, to avoid repetition, a general notation will be used. The general set is given by D, where D should be replaced by B or C depending on which of the two sets is being considered. The data in this general set are given by $\mathbf{w}$, with the i-th sample from the m_D samples in this general set given by

$\mathbf{w}_i = (w_{i1}, \ldots, w_{in_{D_i}})$. The letter $\mathbf{w}$ should be replaced by $\mathbf{x}$ or $\mathbf{y}$, when set B or C is being considered, respectively.

3.1 The autoregressive model of order one

Autoregressive models are a common model for autocorrelated data. Chatfield (2004) defines an autoregressive process X_t, for $t \in \mathbb{Z}$ of order (or lag) p as

$$X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + Z_t \tag{3.1}$$

where the error terms Z_t are mutually independent and identically distributed with mean zero and constant variance. An autoregressive process is used to model autocorrelated data where each observation has a dependence on the value of previous observations. The autocorrelation parameters $\alpha_1, \ldots, \alpha_p$ govern the strength of this dependence. Evidential data which have a sequential nature, with dependence of observations on previous observations can be modelled with such an autoregressive process. In this thesis, autoregressive processes of lag one will be considered. For larger lags, the same methods can be used, but a larger number of parameters will need to be estimated. An assumption will be made that the error terms are Normally distributed. Again, this assumption can be dispensed with if another distribution would better suit the data. This change would necessitate an adjustment to the likelihood function used to estimate the parameters in Section 3.1.2.

There are two models to fit: one for samples in set C and one for samples in set B. The form of the model is the same in both cases, only the parameters are different, so notation corresponding to the general sample, from set D, will be used.

An autoregressive model of lag one ($AR(1)$) is fitted to each sample $\mathbf{w}_i$, so that the relationship between the observations is as follows:

$$w_{it} - \mu = \alpha\,(w_{i,t-1} - \mu) + \epsilon_{it}, \tag{3.2}$$

where $t = 2, \ldots, n_{D_i}$; $\epsilon_{it} \sim N(0, \sigma^2)$ and $w_{i1} \sim N(\mu, \sigma^2)$. The parameter μ is the mean of the process, σ^2 is the variance of the error terms and α is the autocorrelation. A reparametrisation of the specification in (3.1) is used, along with the addition of an intercept, to allow for the modelling of the mean, μ. This model is fitted to each sample i, for $i \in \{1, \ldots, m_D\}$, to obtain parameter estimates for μ, σ^2 and α for each sample in set D.

3.1.1 Prior distributions

A Bayesian approach is used to fit the model, with prior distributions for each of the parameters. The training data are used in conjunction with these prior distributions to obtain draws from the posterior distributions of the parameters of the autoregressive model conditional on the i-th sample.

The prior distributions used for the mean and variance parameters are similar to those used in Rydén (2008) and Richardson and Green (1997), and a truncated Normal prior is used for the autocorrelation parameters, as used in Albert and Chib (1993). The parameter σ^2 is given a hyperparameter, β, which has a hyperprior. Denote all of the parameters by $\theta = (\mu, \sigma^2, \alpha, \beta)$ for brevity. These prior distributions are used to provide compatibility with the hidden Markov model.

The prior distributions used for the parameters in θ are given by

- $\mu \sim \mathrm{N}(\mu_0, V_\mu)$.
- $\sigma^2 \sim \mathrm{IG}(\gamma, \beta)$, where IG denotes the inverse gamma distribution and β is a hyperparameter.
- $\beta \sim \Gamma(g, h)$.
- $\alpha \sim \mathrm{N}(\alpha_0, V_\alpha)$, with the autocorrelation restricted to lie between -1 and 1.

where

- $\mathrm{IG}(\gamma, \beta)$ denotes the inverse gamma distribution such that if $\sigma^2 \sim \mathrm{IG}(\gamma, \beta)$, then

$$f(\sigma^2 \mid \gamma, \beta) = \frac{\beta^\gamma}{\Gamma(\gamma)} (\sigma^2)^{-(\gamma+1)} e^{-\beta/\sigma^2};\ \beta > 0, \gamma > 0, \sigma > 0,$$

- and $\Gamma(g, h)$ denotes the gamma distribution such that if $\beta \sim \Gamma(g, h)$, then

$$f(\beta \mid g, h) = \frac{h^g}{\Gamma(g)} \beta^{g-1} e^{-h\beta};\ g > 0, h > 0, \beta > 0.$$

The specific parameters chosen for these prior distributions for the data relating to traces of cocaine on banknotes can be seen in Section 6.2.1.

3.1.2 Posterior distributions

In order to calculate the likelihood ratio for a questioned sample, an estimate of the between sample distribution of the parameter θ is required, for each of the two sets of training data, B and C. The method for obtaining draws from the posterior distribution of θ, conditional on a single sample of data from one of the two training sets B or C is given in this section. The method for obtaining an estimate of the distribution of the overall parameter θ, conditional on the entire training data set $\mathbf{x}$ or $\mathbf{y}$, using these individual posterior distributions is given in Section 4.1. The model specification in (3.2) can be used to obtain the likelihood of θ for any given sample. Combining this with the density functions associated with the prior distributions given in Section 3.1.1, and using Bayes' theorem, the posterior density function of the parameters θ, given the general sample of data $\mathbf{w}_i$ (substitute for either $\mathbf{x}_i$ or $\mathbf{y}_i$ as appropriate), is given by

$$f(\theta \mid \mathbf{w}_i) \propto f(\mathbf{w}_i \mid \theta) f(\theta). \tag{3.3}$$

Two methods for obtaining draws from this posterior density function $f(\theta \mid \mathbf{w}_i)$ are discussed, a Gibbs sampler and a random walk Metropolis-Hastings sampler.

The prior distributions on the parameters μ, σ^2, α and β are such that the likelihood combined with the priors gives standard distributions for each of the parameters, conditional on the data and the remaining parameters. These standard distributions can be sampled from, and hence a Gibbs sampler can be used to obtain draws from the posterior distribution of θ. The forms of the conditional distributions required and a discussion of the Gibbs sampler are given in Appendix A.1. In practice, it was found that the Gibbs sampler sometimes became trapped in local modes; a Metropolis-Hastings sampler was developed which did not have this issue. This was mainly a problem for the more complicated hidden Markov model, but for consistency (and ease of extension when programming the samplers for both models), the Metropolis-Hastings sampler was used to sample from the $AR(1)$ posterior distributions in the example in Chapters 5 and 6.

The Metropolis-Hastings sampler is a Markov Chain Monte Carlo algorithm, first used in Metropolis et al. (1953) and developed further in Hastings (1970). A summary can be found in Chib and Greenberg (1995). The sampler allows draws to be obtained from the posterior distribution of a parameter (in our case θ) conditional on some data (in our case $\mathbf{w}_i$), $f(\theta \mid \mathbf{w}_i)$. To obtain these draws, it is not necessary to know the normalizing constant of the density function $f(\mathbf{w}_i \mid \theta)$, although in the case of the autoregressive model, this constant is known. The steps for obtaining a chain of draws from $f(\theta \mid \mathbf{w}_i)$ using a generic Metropolis-Hastings algorithm are as follows:

- 1. Initialize chain by setting $\theta^{(0)}$ to some initial values.
- 2. Draw a proposed value θ' from some proposal distribution $q(\theta' \mid \theta^{(0)})$.
- 3. Accept this proposed value with probability A, where

$$A = \frac{f(\theta')f(\mathbf{w}_i \mid \theta')q(\theta^{(0)} \mid \theta')}{f(\theta^{(0)})f(\mathbf{w}_i \mid \theta^{(0)})q(\theta' \mid \theta^{(0)})}.$$

  Here, $f(\theta)$ is the prior density function of the parameter θ and $f(\mathbf{w}_i \mid \theta)$ is the likelihood of θ.
- 4. If θ' was accepted, set $\theta^{(1)} = \theta'$, otherwise set $\theta^{(1)} = \theta^{(0)}$.
- 5. Repeat steps 2-4, replacing 0 by r and 1 by $r+1$, for $r = 2, 3, \ldots, N$, for the desired number of draws N.

The first part of the chain should be discarded to allow the chain to move away from the starting values (known as burn-in).

For the rest of this section, a Metropolis-Hastings algorithm used to obtain draws from $f(\theta \mid \mathbf{w}_i)$, specific to the autoregressive model of order one defined in (3.2) will be discussed.

The probability density function $f(\mathbf{w}_i \mid \theta)$ is given by

$$\begin{aligned} f(\mathbf{w}_i \mid \theta) =& (2\pi\sigma^2)^{-\frac{n_{D_i}}{2}} \exp\left[-\frac{1}{2\sigma^2}(w_{i1}-\mu)^2\right] \\ & \times \exp\left[-\sum_{t=2}^{n_{D_i}}\left(\frac{1}{2\sigma^2}(w_{it}-\mu+\alpha\mu-\alpha w_{i,t-1})^2\right)\right]. \end{aligned} \tag{3.4}$$

The prior distributions of μ, σ^2, α and β were taken to be Normal, inverse gamma, truncated Normal and gamma respectively, with parameters as given in Section 3.1.1, so the joint prior density function, $f(\theta)$, is given by

$$\begin{aligned} f(\theta) =& f(\mu) f(\sigma^2 \mid \beta) f(\beta) f(\alpha) \\ & \propto \exp\left[-\frac{1}{2V_\mu}(\mu-\mu_0)^2\right] \\ & \times \beta^\gamma \sigma^{-(2\gamma+2)} \exp\left[-\frac{\beta}{\sigma^2}\right] \beta^{g-1} \exp(-h\beta) \\ & \times \exp\left[-\frac{1}{2V_\alpha}(\alpha-\alpha_0)^2\right] I(|\alpha|<1), \end{aligned} \tag{3.5}$$

where $I(|\alpha|<1)$ is the indicator function such that

$$\begin{aligned} I(|\alpha|<1) &= 1 \text{ if } |\alpha|<1, \\ &= 0 \text{ if } |\alpha| \geq 1. \end{aligned}$$

Letting the parameters at the r-th step of the Metropolis-Hastings sampler be denoted by $\theta^{(r)} = (\mu^{(r)}, \sigma^{2(r)}, \alpha^{(r)}, \beta^{(r)})$ and the proposed parameters by $\theta' = (\mu', \sigma'^2, \alpha', \beta')$, the Metropolis-Hastings sampler updates the parameters as

$$\begin{aligned} \mu' &= \mu^{(r)} + \varepsilon_1 \\ \log(\sigma'^2) &= \log(\sigma^{2(r)}) + \varepsilon_2 \\ \alpha' &= \alpha^{(r)} + \varepsilon_3 \\ \log(\beta') &= \log(\beta^{(r)}) + \varepsilon_4. \end{aligned}$$

Here, ε_k is a Normally distributed random variable, with zero mean and variance V_k for $k \in \{1,2,3,4\}$. It has been shown (Gelman, Roberts et al. (1996)) that the V_k should be chosen so that the proportion of accepted updates is close to 25%. The values of V_k will vary, depending on the problem, and so should be adjusted so that the proportion of accepted updates is close to this value.

The acceptance probability $A(\theta^{(r)}, \theta' \mid \mathbf{w}_i)$ is given by

$$A(\theta^{(r)}, \theta' \mid \mathbf{w}_i) = \frac{f(\mathbf{w}_i \mid \theta') f(\theta') \sigma'^2 \beta'}{f(\mathbf{w}_i \mid \theta^{(r)}) f(\theta^{(r)}) \sigma^{2(r)} \beta^{(r)}}. \tag{3.6}$$

A random variable U is drawn from a uniform distribution on the interval $[0,1]$, and the updated

parameter θ' is accepted (so $\theta^{(r+1)}$ is set to θ') if $U < \min(1, A(\theta^{(r)}, \theta' \mid \mathbf{w}_i))$. If θ' is not accepted then $\theta^{(r+1)}$ is set to $\theta^{(r)}$.

The terms $\sigma'^2, \beta', \sigma^{2(r)}$ and $\beta^{(r)}$ in (3.6) are included to allow for the fact that the parameters σ^2 and β are updated via their logarithms. The prior density function must therefore be transformed, and these extra terms are the Jacobian of that transformation. The proposal distribution is not included in this expression because it is a multivariate Normal distribution, which is symmetric. As a result, the terms relating to the proposal distributions in the numerator and denominator cancel.

The procedure of proposing and accepting or rejecting new draws, given above, is repeated N times, to acquire N draws $\theta^{(r)}$ for $r \in \{1, \ldots, N\}$ from the posterior density function $f(\theta \mid \mathbf{w}_i)$. This procedure should be used separately for each sample of data $\mathbf{x}_i$ in $\mathbf{x}$ and each sample of data $\mathbf{y}_i$ in $\mathbf{y}$. The parameters associated with sample $\mathbf{x}_i$ are denoted by $\theta^B_{A_i} = (\mu^B_i, (\sigma^B_i)^2, \alpha^B_i, \beta^B_i)$ and the parameters associated with sample $\mathbf{y}_i$ are denoted by $\theta^C_{A_i} = (\mu^C_i, (\sigma^C_i)^2, \alpha^C_i, \beta^C_i)$. The subscript A indicates that these are the parameters for the autoregressive model.

3.2 The autoregressive model with random effects

The model defined in Section 3.1 can be thought of as a fixed effects model. Samples from the posterior distributions of the parameters $\theta^B_{A_i}$ for $i \in \{1, \ldots, m_B\}$ and $\theta^C_{A_i}$ for $i \in \{1, \ldots, m_C\}$, conditional on a single sample of data, are obtained. There is one posterior distribution for each sample in each of the two training sets. A random effects model can also be defined so that, as in (3.2), the relationship between the observations in the general sample $\mathbf{w}_i$ is specified by

$$w_{it} - \mu_i = \alpha_i \, (w_{i,t-1} - \mu_i) + \epsilon_{it}$$

where $t = 2, \ldots, n_{D_i}$; $\epsilon_{it} \sim N(0, \sigma^2_i)$ and $w_{i1} \sim N(\mu_i, \sigma^2_i)$. The dependence of the parameters μ_i, σ^2_i and α_i on the i-th sample has been made explicit using the subscript i.

Now, instead of estimating individual parameters $(\mu_i, \sigma^2_i, \alpha_i)$ for each i (as in Section 3.1), assume that the means μ_i are Normally distributed with

$$\mu_i \sim \mathrm{N}(\mu_\mu, \sigma^2_\mu), \tag{3.7}$$

the variances σ^2_i are distributed as an inverse gamma so that

$$\sigma^2_i \sim \mathrm{IG}(\gamma_V, \beta_V), \tag{3.8}$$

and the autocorrelation parameters α_i are distributed Normally, but restricted to lie between -1 and 1 so that

$$\alpha_i \sim \mathrm{N}(\mu_\alpha, \sigma^2_\alpha) I(|\alpha| < 1). \tag{3.9}$$

Denote the parameters $(\mu_\mu, \sigma_\mu, \gamma_V, \beta_V, \mu_\alpha, \sigma_\alpha)$ by θ. Estimates of θ for each of the two training

data sets are required to evaluate the likelihood ratio for a questioned sample **z**. These estimates are obtained in the form of draws from the posterior distribution of θ, conditional on the entire data set **x** or the entire data set **y**.

3.2.1 Prior distributions

The prior distributions for the parameters in θ are given by

- $\mu_\mu \sim \mathrm{N}(\mu_{\mu_0}, V_{\mu_0})$
- $\sigma_\mu \sim \mathrm{U}(\sigma_{\mu_0}, \sigma_{\mu_1})$
- $\gamma_V \sim \Gamma(\gamma_{\gamma_0}, \beta_{\gamma_0})$
- $\beta_V \sim \Gamma(\gamma_{\beta_0}, \beta_{\beta_0})$
- $\mu_\alpha \sim \mathrm{N}(\mu_{\alpha_0}, V_{\alpha_0})$
- $\sigma_\alpha \sim \mathrm{U}(\sigma_{\alpha_0}, \sigma_{\alpha_1})$

The uniform prior distributions for the standard deviations σ_μ and σ_α were chosen based on recommendations in Gelman (2006).

3.2.2 Posterior distributions

As in Section 3.1.2, draws from the posterior distribution of θ are obtained. This can be done by combining the prior distributions given in Section 3.2.1 with the likelihood of θ and the distributions of the parameters μ_i, σ_i^2 and α_i and then using a random walk Metropolis-Hastings sampler. For the random effects model, the entire data set **w** is used in one Metropolis-Hastings sampler to obtain draws from $f(\theta \mid \mathbf{w})$, rather than having a Metropolis-Hastings sampler for each sample $\mathbf{w}_i$ in **w** (as was done in Section 3.1.2).

To implement a random walk Metropolis-Hastings sampler the parameters θ, μ_i, σ_i^2 and α_i (for $i \in \{1, \ldots, m_D\}$) should be updated using a multivariate Normal proposal distribution. Denote the updated parameters by θ', μ_i', $\sigma_i^{2'}$ and α_i' and the parameters at the r-th step of the sampler by $\theta^{(r)}$, $\mu_i^{(r)}$, $\sigma_i^{2(r)}$ and $\alpha_i^{(r)}$. The acceptance probability for the proposed parameters at the $(r+1)$-th step of the sampler is given by

$$A(\theta^{(r)}, \theta' \mid \mathbf{w}) = \frac{\left(\prod_{i=1}^{m_D} f(\mathbf{w}_i \mid \mu_i', \sigma_i^{2'}, \alpha_i') f(\mu_i', \sigma_i^{2'}, \alpha_i' \mid \theta')\right) f(\theta')}{\left(\prod_{i=1}^{m_D} f(\mathbf{w}_i \mid \mu_i^{(r)}, \sigma_i^{2(r)}, \alpha_i^{(r)}) f(\mu_i^{(r)}, \sigma_i^{2(r)}, \alpha_i^{(r)} \mid \theta^{(r)})\right) f(\theta^{(r)})}, \tag{3.10}$$

where $f(\mathbf{w}_i \mid \mu_i, \sigma_i^2, \alpha_i)$ is the probability density function of $\mathbf{w}_i$ as defined in Section 3.1.2, $f(\mu_i, \sigma_i^2, \alpha_i \mid \theta)$ is given by the product of the probability density functions associated with the three distributions given in (3.7), (3.8) and (3.9), and $f(\theta)$ is the product of the probability density functions associated with the prior distributions given in Section 3.2.1.

It was found for the data considered in Chapters 5 and 6 that standard software packages could be used to obtain draws from $f(\theta \mid \mathbf{w})$ for the autoregressive model with random effects. This was not the case for the standard autoregressive model (with fixed effects) and the hidden Markov model. The rjags package (Plummer (2013)) was therefore used to obtain the required draws; this meant that it was not necessary to implement code for the above Metropolis-Hastings sampler.

The autoregressive model with random effects is similar to the semi-conjugate model considered in Alberink et al. (2013). However, in Alberink et al. (2013), independence is assumed. Here, autocorrelation between adjacent observations is accounted for with the addition of an autocorrelation parameter.

To calculate likelihood ratios, draws from the posterior distribution of the parameter θ are required for each of the training data sets B and C. Denote the parameter for training data set B by $\theta^B_{A_r} = (\mu^B_\mu, \sigma^B_\mu, \gamma^B_V, \beta^B_V, \mu^B_\alpha, \sigma^B_\alpha)$ and the parameter for training data set C by $\theta^C_{A_r} = (\mu^C_\mu, \sigma^C_\mu, \gamma^C_V, \beta^C_V, \mu^C_\alpha, \sigma^C_\alpha)$. The subscript A_r signifies that the autoregressive model with random effects has been used. In general, when the autoregressive model is discussed, the model being referred to is that given in Section 3.1. The model given in this section, using random effects, is included for comparison.

3.3 The hidden Markov model

3.3.1 Introduction

In a hidden Markov model (HMM), each observed data point is associated with an unobserved state. The states form a Markov chain and determine the probability density function of the data point. As an example, consider a set of n observed data points $w_1, w_2, w_3, \ldots, w_n$ (such as inflation rate), with each data point associated with an unobserved binary state $S_1, S_2, S_3, \ldots, S_n$ (such as whether the economy is in boom or bust). Each state S_t for $t \in \{1, 2, \ldots, n\}$ can take a value in $\{0, 1\}$. The probability density function of each w_t can be set with parameters dependent on the associated S_t. For example, the w_t could be taken to be Normally distributed with variance 1 and mean μ_0 (if $S_t = 0$) or μ_1 (if $S_t = 1$) . This model can be represented with the Bayesian network given in figure 3.1.

The states of a hidden Markov model form a Markov chain, which means that the Markov property must hold. This means that conditional on the previous state, each state is independent of all states before that. In other words, if the states S_t, with $t \in \{1, 2, 3, \ldots, n\}$ are a stochastic process taking values in $\{j_1, j_2, \ldots, j_n\}$, then the Markov property holds if for all t,

$$P(S_t = j_t \mid S_1 = j_1, \ldots, S_{t-1} = j_{t-1}) = P(S_t = j_t \mid S_{t-1} = j_{t-1}).$$

An overview of hidden Markov models is given in Rabiner (1989). Scott (2002) discusses the use of MCMC methods for parameter estimation in hidden Markov models.

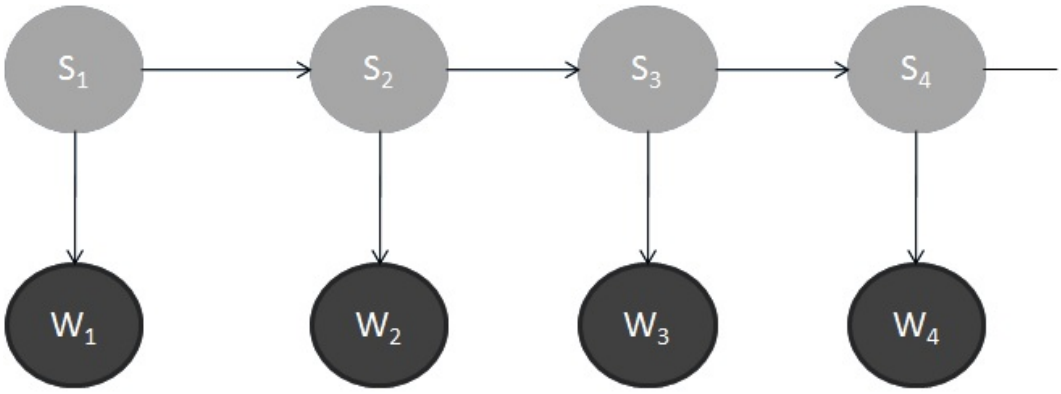

Figure 3.1: Bayesian network of example hidden Markov model, with states S_t and observations w_t for $t \in \{1, \ldots, n\}$

3.3.2 The proposed model

Hidden Markov models can be used to model evidential data which are sequential in time, or spatial in one dimension, and which have an unobserved Markov chain governing the behaviour of the observed data. In this thesis, this Markov chain is assumed to have a finite state space, and to be time-homogeneous, so that the probabilities of moving between states are constant for all t. An extension to the basic model shown in figure 3.1 is used, which allows for dependence of observations on previous observations, as well as between states, so that autocorrelation between adjacent observations is modelled. This means that observations are not independent of one another, conditional on the states: the t-th observation w_t is not independent of previous observations $w_1, \ldots, w_{t-1}$, conditional on the hidden state S_t. A Bayesian network of the proposed model is shown in figure 3.2. These types of models are sometimes known as regime switching models or Markov-switching models (Cappé et al. (2005)); some other examples of Markov switching models can be seen in Section 3.3.3.

The hidden Markov model used in this thesis can be thought of as an extension of the autoregressive model discussed in Section 3.1. The form of the model for each individual observation is as before, but the hidden states allow for multiple sets of parameters, so that parameters can differ between observations, depending on the state. That the states form a Markov chain and are not independent of one another means that dependence between the states can be modelled, unlike in a standard mixture model. For example, with a hidden Markov model it is possible to model a scenario where observations with the same state cluster together.

Four states, corresponding to two different sets of means and variances for the observations, are used. Denote these two sets of means and variances by $a_1 = (\mu_1, \sigma_1^2)$ and $a_2 = (\mu_2, \sigma_2^2)$. These two different sets of parameters are used to model the situation where a sample from B or C contains a mixture of observations, some of which are thought to have means and variances in line with level a_1 and some of which are thought to have means and variances in line with level a_2. The hidden states determine which observations have which level, a_1 or a_2.

A scenario in which this model might be useful would be if level a_2 corresponded to observations which were in some way associated with crime, and level a_1 corresponded to observations which were not associated with crime. Samples in both B and C might be a mixture of such 'crime' and 'not crime' observations. The proportion of 'crime' observations within each sample might be larger for samples in C than in B. In this case the proportion of observations with each level is important and must be modelled in order to discriminate between the two propositions.

The hidden states

The states associated with the two training sets, B and C are given by:

- $\mathbf{S_B} = \{S_{B_{it}}; i = 1,\ldots,m_B, t = 1,\ldots,n_{B_i}\}$: the states for data $\mathbf{x}$, with a state $S_{B_{it}}$ for each observation x_{it}. Each state $S_{B_{it}}$ can take values in $\{1,2,3,4\}$.
- $\mathbf{S_C} = \{S_{C_{it}}; i = 1,\ldots,m_C, t = 1,\ldots,n_{C_i}\}$: the states for data $\mathbf{y}$, with a state $S_{C_{it}}$ for each observation y_{it}. Each state $S_{C_{it}}$ can take values in $\{1,2,3,4\}$.

The states associated with the questioned sample $\mathbf{z}$ are given by:

- $\mathbf{R} = \{R_t; t = 1,\ldots,n\}$: the states for the questioned sample $\mathbf{z}$, with one state for each of the n observations. Each state R_t can take values in $\{1,2,3,4\}$.

As before, a general notation is used. The states associated with the general sample $\mathbf{w}_i$ are given by $\mathbf{S}_i = (S_{i1},\ldots,S_{in_{D_i}})$.

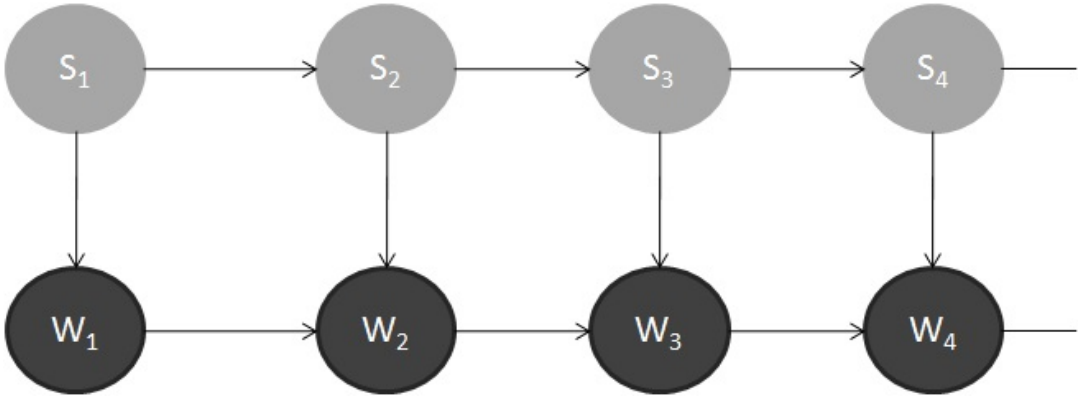

Figure 3.2: Bayesian network of the proposed hidden Markov model, with states S_t and observations w_t for $t \in \{1,\ldots,n\}$

The two possible sets of mean and variance levels are a_1 and a_2. For identifiability and without loss of generality, let the mean of a_1 be smaller than that of a_2. The model has four hidden states, defined by which of a_1 or a_2 is associated with the current and previous observations. These states are given by

State (S)	Previous data point	Current data point
1	a_1	a_1
2	a_1	a_2
3	a_2	a_1
4	a_2	a_2

(3.11)

The transition matrix, with two transition probabilities p_{01} and p_{10}, which gives the probabilities of moving between these states, is

$$\mathbf{P} = \begin{pmatrix} 1-p_{01} & p_{01} & 0 & 0 \\ 0 & 0 & p_{10} & 1-p_{10} \\ 1-p_{01} & p_{01} & 0 & 0 \\ 0 & 0 & p_{10} & 1-p_{10} \end{pmatrix}. \tag{3.12}$$

The initial distribution of the hidden state of the first observation is given by

$$\pi_{INIT} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}.$$

Four states are used, so that the parameters of both the current and previous observations are described in each state. This is necessary because the probability density function of each observation, conditional on the previous observation, depends on the mean of the previous observation. By using four states instead of two, an extra line of dependence in the Bayesian network of figure 3.2, between each observation and the previous state, is not required. This results in a transition matrix containing zeros, as it is impossible to pass from, for example, a state with a current observation having mean and variance level a_1 to a state with a previous observation having mean and variance level a_2.

The model

The probability model for sample $\mathbf{w}_i$, conditional on the states $\mathbf{S}_i$ is given by

$$w_{it} - \mu^{(1)}_{S_{it}} = \alpha(w_{i,t-1} - \mu^{(2)}_{S_{it}}) + \epsilon_{S_{it}} \tag{3.13}$$

for

$$t \in \{2,3,\ldots,n_{D_i}\}, \text{ where } \epsilon_{S_{it}} \sim \mathrm{N}(0,\sigma^2_{S_{it}}), \text{ and } w_{i1} \sim \mathrm{N}(\mu^{(1)}_{S_{i1}},\sigma^2_{S_{i1}})$$

and

- The subscript S_{it} indicates that the parameter associated with state S_{it} should be used.
- $\mu^{(2)}_{S_{it}}$ is the mean of the previous observation and hence has the same value as $\mu^{(1)}_{S_{i,t-1}}$.

As can be seen from (3.13), the probability density function of observation w_{it}, conditional on the state S_{it} and the previous observation $w_{i,t-1}$ depends on four parameters: $\mu^{(2)}_{S_{it}}$, $\mu^{(1)}_{S_{it}}$, $\sigma^2_{S_{it}}$ and α. The mean $\mu^{(2)}_{S_{it}}$ is the mean level of the previous observation. The mean $\mu^{(1)}_{S_{it}}$ is the mean level of the current observation. According to the definition of the states given in (3.11), the parameters associated with each state, written in the order $(\mu^{(2)}_{S_{it}}, \mu^{(1)}_{S_{it}}, \sigma^2_{S_{it}}, \alpha)$ must therefore be

$$\begin{aligned}
&\text{State 1}: (\mu_1, \mu_1, \sigma_1^2, \alpha)\\
&\text{State 2}: (\mu_1, \mu_2, \sigma_2^2, \alpha)\\
&\text{State 3}: (\mu_2, \mu_1, \sigma_1^2, \alpha)\\
&\text{State 4}: (\mu_2, \mu_2, \sigma_2^2, \alpha).
\end{aligned}$$

Note that only two parameters are required for each of the mean and variance, as there are only two mean and variance levels. Therefore, $\mu^{(1)}_1 = \mu^{(2)}_1 = \mu^{(2)}_2 = \mu^{(1)}_3 = \mu_1$, $\mu^{(1)}_2 = \mu^{(2)}_3 = \mu^{(1)}_4 = \mu^{(2)}_4 = \mu_2$, $\sigma_1 = \sigma_3$ and $\sigma_2 = \sigma_4$.

The model for each sample is, therefore, specified by the parameters

$$(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \alpha, p_{01}, p_{10}, p_1, p_2, p_3, p_4).$$

3.3.3 Other examples of this type of hidden Markov model in the literature

A hidden Markov model similar to that in Section 3.3.2 was introduced in Hamilton (1989) to model GNP (gross national product). Hidden states were used to represent the state of the business cycle (either expansion or contraction). For a sample $\mathbf{w} = (w_1, \ldots, w_n)$, where w_t is related to the US GNP at time t, the relationship between the observations and hidden states is given in Hamilton (1989) as

$$w_t - \mu_{S_t} = \sum_{r=1}^{p} \alpha_r (w_{t-r} - \mu_{S_{t-r}}) + \epsilon_t$$

for $t \in \{1, \ldots, n\}$, where p is the lag of the time series and $\epsilon_t \sim N(0, \sigma^2)$. The subscript S_t indicates that the parameter is dependent on the hidden state S_t. This model was termed the switching autoregressive process and in Hamilton (1989) it was assumed that there were two hidden states. There are several differences between this model and the model proposed in Section 3.3.2. The main difference is that the variance, σ^2, does not depend on the hidden state. In (3.13), the variance parameters are dependent on the value of the hidden state. In addition, Hamilton (1989) allows for autocorrelation at lags of greater than one whereas the model used in (3.13) allows only for lag one autocorrelation. Hamilton (1989) uses two states instead of four; using four states allows for more efficient Markov Chain Monte Carlo algorithms to be used for parameter estimation because the hidden states can be sampled from in one block, using the method described in Chib (1996). Hamilton (1989) obtains estimates of the parameters and the hidden states using maximum likelihood

methods rather than using Bayesian techniques to obtain draws from the posterior distributions of the parameters, conditional on the data (as will be done in Section 3.3.6).

Cappé et al. (2005) give a more general formulation of a switching autoregressive process as

$$w_t - \mu_{S_t} = \sum_{r=1}^{p} \alpha_{rS_t}(w_{t-r} - \mu_{S_{t-r}}) + \epsilon_t \tag{3.14}$$

where this time $\epsilon_t \sim N(0, \sigma^2_{S_t})$. With this formulation, the autocorrelation, variance and mean parameters all depend on the hidden state. McCulloch and Tsay (1994), Albert and Chib (1993) and Kim and Nelson (1999) apply models of this sort to real data; a discussion of these applications follows.

McCulloch and Tsay (1994) use a model given by

$$w_t = \mu_{S_t} + \sum_{r=1}^{p} \alpha_{rS_t} w_{t-r} + \epsilon_t,$$

for $\epsilon_t \sim N(0, \sigma^2_{S_t})$. As with Hamilton (1989), it is assumed that there are two possible states. The intercept terms (μ_{S_t}), the autocorrelation parameters (α_{rS_t}) and the variances ($\sigma^2_{S_t}$) are all state dependent, as in (3.14). Unlike in (3.14), this model uses an intercept term instead of considering the differences ($w_t - \mu_{S_t}$). This means that the probability density function of the observation w_t, conditional on the current state S_t, and the previous observation w_{t-1} does not depend on the states associated with previous observations. To obtain parameter estimates, a Gibbs sampler is used alongside prior distributions on the parameters, to obtain Bayesian estimates of the posterior distributions of the parameters and hidden states, conditional on the data; this is unlike the maximum likelihood approach used in Hamilton (1989).

Albert and Chib (1993) consider the differences ($w_t - \mu_{S_t}$) rather than having an intercept term, and as such the model in Albert and Chib (1993) is similar to the general model given in (3.14). Means and variances are taken to be state dependent, but the autocorrelation parameters are assumed independent of the state. Two states are used, as in Hamilton (1989) and McCulloch and Tsay (1994). The form of the model in Albert and Chib (1993) is the same as the form of the model proposed in (3.13) (although with two states instead of four). Means (μ_{S_t}) and variances ($\sigma^2_{S_t}$) depend on the state and the autocorrelation parameter (α) is taken to be the same for each state. However, a different approach is used to fit the model. Albert and Chib (1993) use a Gibbs sampler to obtain draws from the posterior distribution of both the parameters and the hidden states, conditional on the data. In Section 3.3.6, a Metropolis-Hastings sampler will be used to obtain draws from the posterior distribution of the parameters, without needing to sample from the hidden states. Integrating out the hidden states in this way is beneficial if information about the hidden states is not required, because correlation between the parameters and the hidden states can be high. High correlation between parameters can affect the convergence of a Markov Chain Monte Carlo sampler.

Another application of a switching autoregressive process can be seen in Kim and Nelson (1999). A model similar to the model in Hamilton (1989) is used, with a mean which depends on the state.

Bayesian approaches using a Gibbs sampler, as in Albert and Chib (1993) and McCulloch and Tsay (1994), are used to obtain draws from the posterior distribution of the parameters and the hidden states. Extensions are given to allow for the mean and variance to shift after a given changepoint.

All of the examples discussed were applied to econometric time series data but autoregressive switching models of this sort have been applied to other types of data. Some examples include their use for the modelling of the evolution of wind speed (Ailliot and Monbet (2012)) and their use for the modelling of disease outbreaks (Lu et al. (2010)).

3.3.4 Prior distributions

As with the autoregressive model, a Bayesian approach is used to fit the hidden Markov model, with prior distributions for each of the parameters. The training data are used in combination with these prior distributions to obtain posterior distributions for the following parameters, which are denoted for brevity by θ:

$$\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \alpha, p_{01}, p_{10}, \beta_1, \beta_2).$$

A truncated Normal prior is used for the autocorrelation parameter, as in Albert and Chib (1993). One of the hyperparameters of each of the two variances is given a hyperprior, as in Richardson and Green (1997) and Rydén (2008). These hyperparameters are denoted β_1 (hyperparameter for σ_1^2) and β_2 (hyperparameter for σ_2^2). Examples in Richardson and Green (1997) show that using a hyperprior on the parameters β_1 and β_2 in mixture models reduces the sensitivity of the posterior distribution to the parameters of the prior distributions chosen. Prior distributions are not used for the parameters associated with the initial distribution of the hidden states, p_1, p_2, p_3 and p_4. Instead, as discussed on p. 478 of Cappé et al. (2005), these parameters are taken to be known, setting $p_1 = p_2 = p_3 = p_4 = 0.25$. As only one hidden state is associated with this initial distribution it would not be possible to obtain good estimates of these parameters, were they taken to be unknown.

It was found that the posterior distribution could be sensitive to the hyperparameters used in the prior distributions of the transition probabilities, especially when small transition probabilities were permitted and the sample size was small. This issue is discussed in Gassiat and Rousseau (2013), where examples are given of the unstable behaviour of a Metropolis-Hastings sampler when the hyperparameters of a beta prior distribution on the transition probabilities of a hidden Markov model are small and the model is overfitted (with too many hidden states). Gassiat and Rousseau (2013) recommend that large hyperparameters are used with a Dirichlet prior distribution (which is the generalisation of a beta distribution) when the number of hidden states is unknown.

If the hyperparameters of a beta prior distribution are large, then the prior assigns less weight to very small values of transition probabilities. Small values for the transition probabilities would imply that the states are very 'sticky', e.g. that once in state one or four, the probability of leaving the state is very small. For examples tested, it was found that when very small transition probabilities were allowed, the sampler would sometimes allocate all observations either to state one or to state

four, resulting in unstable estimates of the posterior distributions of the parameters for the states with which no observations were associated. If all of the observations have the same state then the model fitted is essentially the autoregressive model, not the hidden Markov model, implying that the problem is that the model has been overspecified, and the extension to the hidden Markov model is not necessary. As discussed in Gassiat and Rousseau (2013), if less weight is given to small transition probabilities, the sampler merges the states by setting the parameters associated with each state to be equal, instead of allocating all observations to one state, if the model is overspecified. If the states are merged then the estimated posterior distributions of the means and variances associated with different states will be the same and therefore there are likely to be observations assigned to each state. This resolves the problem of unstable estimates of the posterior distributions, because there are observations available for the estimation of the posterior distributions of each of the parameters. Checking which of the two models (autoregressive or hidden Markov) best fits the data can then be carried out separately. See Chapter 6 for an example which uses Bayes Factors for model selection. To mitigate some of the issues discussed, a truncated beta distribution is used as the prior distribution for the transition probabilities. A truncated beta distribution puts no weight at all on very small values of transition probabilities.

The prior distributions for the hidden Markov model are given by

- $\mu_1, \mu_2 \sim \mathrm{N}(\mu_0, V_\mu)$.
- $\sigma_1^2 \sim \mathrm{IG}(\gamma, \beta_1)$.
- $\sigma_2^2 \sim \mathrm{IG}(\gamma, \beta_2)$.
- $\beta_1, \beta_2 \sim \Gamma(g, h)$.
- $\alpha \sim \mathrm{N}(\alpha_0, V_\alpha)$, with the autocorrelation restricted to lie between -1 and 1.
- $p_{01}, p_{10} \sim \mathrm{Beta}(a, b)$, and restricted to lie between $2/n$ and $(n-2)/n$, where n is the number of observations in the sample.

The choice of the value of the hyperparameters for these prior distributions can be made subjectively if an appropriate expert is available. Alternatively, values can be chosen which are relatively uninformative, with the exception of the hyperparameters for the transition probabilities, which for the reasons discussed above, might need to be set to larger values if the sampler produces unstable estimates. The hyperparameters chosen for the prior distributions for the data relating to traces of cocaine on banknotes are given in Section 6.2.3.

3.3.5 Likelihood

There are two likelihood functions of interest for the hidden Markov model. One is the joint likelihood function of the parameters θ and the hidden states for the i-th general sample, $\mathbf{S}_i$, given by $L(\theta, \mathbf{S}_i) =$

$f(\mathbf{w}_i \mid \theta, \mathbf{S}_i)$. The other likelihood function of interest is the marginal likelihood of the parameters θ, given by $L(\theta) = f(\mathbf{w}_i \mid \theta)$. The joint likelihood of the parameters and the hidden states will be used in Section 3.3.6 to develop a Gibbs sampler. The marginal likelihood $L(\theta)$ will be used to develop a Metropolis-Hastings sampler. The joint likelihood $L(\theta, \mathbf{S}_i)$ is easily computed (see Appendix A.2) using (3.13) given in Section 3.3.2. The marginal likelihood $L(\theta)$ requires the summation of $f(\mathbf{w}_i, \mathbf{S}_i \mid \theta)$ over all possible state sequences. As the number of observations increases, this quickly becomes impossible. The forward-algorithm, as described in Rabiner (1989) and Scott (2002) exploits the conditional structure of a hidden Markov model to sum the hidden states out of $f(\mathbf{w}_i, \mathbf{S}_i \mid \theta)$ in far fewer steps. In this section, this forward-algorithm is described. For brevity, θ is taken to denote the parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \alpha, p_{01}, p_{10}, \beta_1, \beta_2, p_1, p_2, p_3, p_4)$, so including the initial distribution of the first state, unlike the standard definition of θ used elsewhere in this thesis. A straightforward adjustment to the algorithm in Rabiner (1989) is needed to allow for the dependence of w_{it} on $w_{i,t-1}$; this adjustment is given below and can also be seen on p. 332 of Frühwirth-Schnatter (2006).

The required likelihood is given by:

$$f(\mathbf{w}_i \mid \theta) = \sum_{\mathbf{S}_i} f(\mathbf{w}_i, \mathbf{S}_i \mid \theta).$$

As described in Rabiner (1989) and Scott (2002), this likelihood is calculated by recursively calculating the forward variables $\beta_t(s)$, defined as

$$\beta_t(s) = f(w_{i1}, w_{i2}, \ldots, w_{it}, S_{it} = s, \mid \theta).$$

The forward variable $\beta_t(s)$ can be written as

$$\begin{aligned}
\beta_t(s) &= \sum_{k=1}^{4} f(w_{i1}, w_{i2}, \ldots, w_{it}, S_{it} = s, S_{i,t-1} = k \mid \theta) \\
&= \sum_{k=1}^{4} f(w_{it}, S_{it} = s \mid \theta, w_{i1}, w_{i2}, \ldots, w_{i,t-1}, S_{i,t-1} = k) f(w_{i1}, w_{i2}, \ldots, w_{i,t-1}, S_{i,t-1} = k \mid \theta) \\
&= \sum_{k=1}^{4} f(w_{it}, S_{it} = s \mid \theta, w_{i1}, w_{i2}, \ldots, w_{i,t-1}, S_{i,t-1} = k) \beta_{t-1}(k) \\
&= \sum_{k=1}^{4} f(w_{it} \mid \theta, w_{i1}, w_{i2}, \ldots, w_{i,t-1}, S_{i,t-1} = k, S_{it} = s) \\
&\times f(S_{it} = s \mid \theta, w_{i1}, w_{i2}, \ldots, w_{i,t-1}, S_{i,t-1} = k) \beta_{t-1}(k) \\
&= f(w_{it} \mid \theta, w_{i,t-1}, S_{it} = s) \sum_{k=1}^{4} f(S_{it} = s \mid \theta, S_{i,t-1} = k) \beta_{t-1}(k). \qquad (3.15)
\end{aligned}$$

The first term in the last step is simplified because, conditional on the current state S_{it} and the previous observation $w_{i,t-1}$, the observation w_{it} is independent of previous states and previous observations. The second term in the last step is simplified because, conditional on the previous state $S_{i,t-1}$, the state S_{it} is independent of the previous observations $w_{i1}, \ldots, w_{i,t-1}$, because the value w_{it} is not conditioned upon. This is because, considering the Bayesian network given in figure 3.2,

the nodes S_{it} and $w_{i,t-1},\ldots,w_{i,1}$ are d-separated when $S_{i,t-1}$ is known (diverging connection) and w_{it} is not known (converging connection). For more information on this, see p. 41 of Taroni, Aitken et al. (2006). That this simplification still occurs, even though w_{it} is dependent on $w_{i,t-1}$, unlike in a standard hidden Markov model, is key in the extension of the forward algorithm to allow for dependence of w_{it} on $w_{i,t-1}$.

In (3.15), the term $f(w_{it} \mid \theta, w_{i,t-1}, S_{it} = s)$ is Normally distributed, with mean $\mu_s^{(1)} + \alpha(w_{i,t-1} - \mu_s^{(2)})$ and variance σ_s^2. The mean and variance parameters depend on the value of the hidden state S_{it}. The term $f(S_{it} = s \mid \theta, S_{i,t-1} = k)$ is given by the transition matrix of the hidden states.

The initial forward variable $\beta_1(s)$ can be calculated using the initial distributions of w_{i1} and S_{i1} and the relationship $\beta_1(s) = f(w_{i1}, S_{i1} \mid \theta) = f(w_{i1} \mid \theta, S_{i1} = s) f(S_{i1} = s \mid \theta)$. From this, the remaining forward variables can be calculated recursively. The required likelihood is then given by:

$$f(\mathbf{w}_i \mid \theta) = \sum_{s=1}^{4} \beta_n(s).$$

3.3.6 Posterior distributions

In this section, the methods discussed in Section 3.1.2 for estimating the posterior distribution of θ conditional on the data $\mathbf{w}_i$ are extended for use with the hidden Markov model. The parameters to be estimated for this model are $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \alpha, p_{01}, p_{10}, \beta_1, \beta_2)$.

Gibbs sampler

The prior distributions chosen for the hidden Markov model are such that the distributions of each of the parameters, conditional on all other parameters, the data, and the hidden states, are standard distributions which can be sampled from. Chib (1996) gives an algorithm for sampling from the joint distribution of all of the hidden states, conditional on the data and the parameters. An alternative is to sample each state in turn, conditional on the data, the parameters, and all of the other states (an example can be seen in Albert and Chib (1993)). As discussed in Scott (2002), sampling from all of the hidden states in one block should improve the mixing of the algorithm, so this is the preferred method here. As all of the necessary conditional distributions can be sampled from, it is possible to use a Gibbs sampler to obtain draws from the posterior density function of the parameters and the hidden states given the data, $f(\theta, \mathbf{S}_i \mid \mathbf{w}_i)$. From this, draws from the marginal density function $f(\theta \mid \mathbf{w}_i)$ can be obtained. A description of the Gibbs sampler for this problem, including the method for sampling the hidden states, is given in Appendix A.2, along with the forms of the conditional distributions required to implement it.

As discussed in Section 3.1.2, the Gibbs sampler did not perform well when it was used to obtain draws from the posterior distribution of the parameters of the the hidden Markov model for the data relating to traces of cocaine on banknotes (Chapters 5 and 6). For some samples, very different posterior distributions were obtained when the sampler was initialised from different starting points.

The Gibbs sampler is, however, easier to implement than the Metropolis-Hastings sampler, because it does not require a good proposal distribution to be found. There are examples in the literature of Gibbs samplers performing well with similar hidden Markov models (McCulloch and Tsay (1994); Albert and Chib (1993); Kim and Nelson (1999)). It is recommended that if the Gibbs sampler is used, convergence is checked carefully; this can be done by starting several chains from different and overdispersed starting points (relative to the posterior distribution) and comparing the results.

Metropolis-Hastings sampler

The questioned sample **z** has its own set of hidden states, **R**. It is expected that if, for example, proposition H_C is true, and the questioned sample is from the set C, then the transition probabilities associated with the hidden states **R** are from the same distribution as the transition probabilities associated with the hidden states of samples in C. However, the hidden states themselves, given the transition probabilities, are independent of the hidden states of samples in C. It is therefore not necessary to estimate the hidden states for a sample associated with H_C or H_B for use in calculating the likelihood ratio, because there is no expectation of a relationship between the hidden states of the questioned sample and the hidden states of the training data, other than through the transition probabilities. In this section, a Metropolis-Hastings sampler is developed that draws directly from the posterior distribution associated with the density function $f(\theta \mid \mathbf{w}_i)$, avoiding sampling from the hidden states. Drawing directly from $f(\theta \mid \mathbf{w}_i)$ in this way should improve the mixing of the sampler by decreasing autocorrelation between draws from the sampler, because the hidden states and the parameters are likely to be highly correlated (Liu (1994)). Sampling directly from $f(\theta \mid \mathbf{w}_i)$ is not possible with a Gibbs sampler, as the full set of conditional distributions cannot be fully specified, with normalising constants.

The Metropolis-Hastings sampler draws from the posterior distribution of θ, conditional on a single sample, and has the same general form as the sampler given in Section 3.1.2. The required posterior density function $f(\theta \mid \mathbf{w}_i)$ is proportional to the product of the likelihood, $f(\mathbf{w}_i \mid \theta)$ and the prior density function $f(\theta)$. To obtain this likelihood $f(\mathbf{w}_i \mid \theta)$, the hidden states must be summed out of the likelihood $f(\mathbf{w}_i, \mathbf{S}_i \mid \theta)$. The algorithm used for this can be seen in Section 3.3.5.

Taking the prior distributions from Section 3.3.4, the parameters μ_1,μ_2, σ_1^2,σ_2^2, α, p_{01}, p_{10}, β_1, β_2

have the following joint prior density function, $f(\theta)$:

$$\begin{aligned}
f(\theta) &= f(\mu_1)f(\mu_2)f(\sigma_1^2 \mid \beta_1)f(\sigma_2^2 \mid \beta_2)f(\beta_1)f(\beta_2)f(\alpha)f(p_{01})f(p_{10}) \\
&\propto \exp\left[-\frac{1}{2V_\mu}(\mu_1-\mu_0)^2\right] \times \exp\left[-\frac{1}{2V_\mu}(\mu_2-\mu_0)^2\right] \\
&\times \beta_1^\gamma \sigma_1^{-(2\gamma+2)} \exp\left[-\frac{\beta_1}{\sigma_1^2}\right] \beta_1^{g-1}\exp(-h\beta_1) \times \beta_2^\gamma \sigma_2^{-(2\gamma+2)} \exp\left[-\frac{\beta_2}{\sigma_2^2}\right] \beta_2^{g-1}\exp(-h\beta_2) \\
&\times \exp\left[-\frac{1}{2V_\alpha}(\alpha-\alpha_0)^2\right] I(|\alpha|<1) \\
&\times p_{01}^{a-1}(1-p_{01})^{b-1} p_{10}^{a-1}(1-p_{10})^{b-1} \mathrm{I}\left(\frac{2}{n_{D_i}} < p_{01} < \frac{(n_{D_i}-2)}{n_{D_i}}\right) \mathrm{I}\left(\frac{2}{n_{D_i}} < p_{10} < \frac{(n_{D_i}-2)}{n_{D_i}}\right). \quad (3.16)
\end{aligned}$$

If the parameters at step r of the Metropolis-Hastings sampler are denoted by

$$\theta^{(r)} = (\mu_1^{(r)}, \mu_2^{(r)}, \sigma_1^{2(r)}, \sigma_2^{2(r)}, \alpha^{(r)}, p_{01}^{(r)}, p_{10}^{(r)}, \beta_1^{(r)}, \beta_2^{(r)})$$

and the proposed parameters are denoted by

$$\theta' = (\mu_1', \mu_2', \sigma_1'^2, \sigma_2'^2, \alpha', p_{01}', p_{10}', \beta_1', \beta_2'),$$

then the Metropolis-Hastings sampler updates the parameters as follows:

$$\begin{aligned}
\mu_1' &= \mu_1^{(r)} + \varepsilon_1 \\
\mu_2' &= \mu_2^{(r)} + \varepsilon_2 \\
\log(\sigma_1'^2) &= \log(\sigma_1^{2(r)}) + \varepsilon_3 \\
\log(\sigma_2'^2) &= \log(\sigma_2^{2(r)}) + \varepsilon_4 \\
\alpha' &= \alpha^{(r)} + \varepsilon_5 \\
\log(p_{01}'/(1-p_{01}')) &= \log(p_{01}^{(r)}/(1-p_{01}^{(r)})) + \varepsilon_6 \\
\log(p_{10}'/(1-p_{10}')) &= \log(p_{10}^{(r)}/(1-p_{10}^{(r)})) + \varepsilon_7 \\
\log(\beta_1') &= \log(\beta_1^{(r)}) + \varepsilon_8 \\
\log(\beta_2') &= \log(\beta_2^{(r)}) + \varepsilon_9.
\end{aligned}$$

Here, ε_k is a Normally distributed random variable, with zero mean and variance V_k for $k \in \{1,2,\ldots,9\}$. As before, the V_k should be chosen so that the number of accepted updates is close to 25% (Gelman, Roberts et al. (1996)).

The acceptance probability $A(\theta^{(r)}, \theta' \mid \mathbf{w}_i)$ is given by

$$A(\theta^{(r)}, \theta' \mid \mathbf{w}_i) = \frac{f(\mathbf{w}_i \mid \theta')f(\theta')\sigma_1'^2\sigma_2'^2\beta_1'\beta_2'(1-p_{01}')p_{01}'(1-p_{10}')p_{10}'}{f(\mathbf{w}_i \mid \theta^{(r)})f(\theta^{(r)})\sigma_1^{2(r)}\sigma_2^{2(r)}\beta_1^{(r)}\beta_2^{(r)}(1-p_{01}^{(r)})p_{01}^{(r)}(1-p_{10}^{(r)})p_{10}^{(r)}}. \quad (3.17)$$

A random variable U is drawn from a uniform distribution on the interval $[0,1]$, and the updated

parameter θ' is accepted (so $\theta^{(r+1)}$ is set to θ') if $U < \min(1, A(\theta^{(r)}, \theta' \mid \mathbf{w}_i))$. If θ' is not accepted, then $\theta^{(r+1)}$ is set to $\theta^{(r)}$.

The term $\sigma_1^{'2}\sigma_2^{'2}\beta_1'\beta_2'(1-p_{01}')p_{01}'(1-p_{10}')p_{10}'$ in the numerator of (3.17) and the term $\sigma_1^{2(r)}\sigma_2^{2(r)}\beta_1^{(r)}\beta_2^{(r)}(1-p_{01}^{(r)})p_{01}^{(r)}(1-p_{10}^{(r)})p_{10}^{(r)}$ in the denominator are the Jacobians of the log transformations of σ_1^2,σ_2^2, β_1 and β_2 and the logistic transformations of p_{01} and p_{10}. The steps required for the calculation of the likelihood $f(\mathbf{w}_i \mid \theta)$ are given in Section 3.3.5. As for the autoregressive model, the proposal distribution is symmetric, and hence is not included in the expression.

As discussed in Scott (2002) and Frühwirth-Schnatter (2001), the likelihood of θ is invariant under certain permutations of the states. In other words, the likelihood is unchanged if the state labels (1,2,3,4) are swapped with (4,3,2,1), with an associated swapping of the labels of the parameters associated with each state (so that e.g. μ_1 is relabelled as μ_2). As such, the marginal posterior density functions of each of the parameters should be bimodal (with the exception of α, which is not state dependent). For example, the marginal posterior density function estimated from the draws from the sampler for μ_1 should be a combination of the true marginal posterior density function of μ_1 (the lower level mean) and the true marginal posterior density function of μ_2 (the upper level mean) if the sampler is mixing well, as the sampler does not constrain $\mu_1 < \mu_2$. Frühwirth-Schnatter (2001) suggests using a permutation sampler to improve the mixing of a Markov chain Monte Carlo sampler where this label switching problem exists. At the end of each run of the sampler, the labels are permuted from (1,2,3,4) to (4,3,2,1) with probability 0.5. This is added as a final step to the algorithm as follows:

- Draw a sample b, from a Bernoulli random variable with parameter $p = 0.5$.
- If $b = 0$, take no action. If $b = 1$, swap $\mu_1^{(r+1)}$ with $\mu_2^{(r+1)}$, $p_{01}^{(r+1)}$ with $p_{10}^{(r+1)}$, $\beta_1^{(r+1)}$ with $\beta_2^{(r+1)}$ and $\sigma_1^{2(r+1)}$ with $\sigma_2^{2(r+1)}$.

The posterior density function of the parameters is used to evaluate the probability density function of measurements on a questioned sample. As this probability density function is also invariant under permutation of the state labels, it is not necessary to constrain the sampler by insisting that $\mu_1 < \mu_2$, which would separate out the bimodal posterior density functions. Instead, the bimodal posterior density functions can be used directly in the calculation of the likelihood for the questioned sample.

The procedure given above is repeated N times, to acquire draws $\theta^{(r)}$, for $r \in \{1, \ldots, N\}$, from the posterior density function $f(\theta \mid \mathbf{w}_i)$. It should be used separately for each sample of data in B and each sample of data in C. The parameters associated with sample $\mathbf{x}_i$ in the training set B are denoted by

$$\theta_{H_i}^B = (\mu_{1_i}^B, \mu_{2_i}^B, (\sigma_{1_i}^B)^2, (\sigma_{2_i}^B)^2, \alpha_{1_i}^B, p_{01_i}^B, p_{10_i}^B, \beta_{1_i}^B, \beta_{2_i}^B)$$

and the parameters associated with sample $\mathbf{y}_i$ in the training set C are denoted by

$$\theta_{H_i}^C = (\mu_{1_i}^C, \mu_{2_i}^C, (\sigma_{1_i}^C)^2, (\sigma_{2_i}^C)^2, \alpha_{1_i}^C, p_{01_i}^C, p_{10_i}^C, \beta_{1_i}^C, \beta_{2_i}^C).$$

The subscript H is used to indicate that these are the parameters for the hidden Markov model.

3.4 The nonparametric model

Nonparametric models make no assumption that the data are drawn from a particular probability distribution. The autoregressive model and the hidden Markov model assume a Normal distribution for the error terms, ϵ_{it}. A nonparametric model dispenses with this assumption. Use of a nonparametric model for the data also means that there is no need for prior distributions for the parameters. The nonparametric method used in this section to estimate density functions from data is kernel density estimation. A kernel density estimate is constructed from some kernel $K(.)$, which is usually a symmetric probability density function, and a bandwidth h with $h > 0$. If x_j for $j \in \{1, \ldots, n\}$ are n observations with underlying probability density function f, then the kernel density estimate of this function f is defined in Silverman (1986) as

$$\hat{f}(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right).$$

An earlier application of kernel density estimation for independent observations in forensic science is given in Aitken and Taroni (2004). In this thesis, data are assumed to be autocorrelated, so conditional kernel density estimates must be used (estimates of conditional density functions). In this section, conditional kernel density estimates are obtained for autocorrelated data of lag one. In the next chapter it is shown how these conditional density estimates can be used to calculate likelihood ratios for autocorrelated forensic data.

The m_D general samples $\mathbf{w}_i$ for $i \in \{1, \ldots, m_D\}$ are used to obtain m_D estimates of the probability density function of an observation from a sample in set D, conditional on the previous observation in the sample. The estimated conditional density function based on the i-th sample in the training set D is denoted by $\hat{f}_{D_i}(\cdot \mid \cdot)$. This estimated conditional density function, evaluated at each pair of observations (z_{t-1}, z_t) in the questioned sample $\mathbf{z} = (z_1, \ldots, z_n)$ for $t \in \{2, \ldots, n\}$, is used to give an estimate of the joint density function of $\mathbf{z}$ of

$$\hat{f}_{D_i}(z_1, z_2, \ldots, z_n) = \hat{f}_{D_i}(z_1)\hat{f}_{D_i}(z_2 \mid z_1) \ldots \hat{f}_{D_i}(z_n \mid z_{n-1}),$$

allowing for autocorrelation of lag one. The estimate of the marginal density function $\hat{f}_{D_i}(z_1)$ is obtained using a univariate kernel density estimate (Silverman (1986); Aitken and Taroni (2004)). The conditional density functions $f_{D_i}(z_t \mid z_{t-1})$ can be estimated nonparametrically, using kernel density estimation, at the point z_t, conditioned on the value of z_{t-1} by

$$\hat{f}_{D_i}(z_t \mid z_{t-1}) = \frac{\hat{g}_{D_i}(z_t, z_{t-1})}{\hat{r}_{D_i}(z_{t-1})}. \tag{3.18}$$

The functions $\hat{g}_{D_i}$ and $\hat{r}_{D_i}$ are kernel density estimates, based on the i-th sample in training set D, given by

$$\hat{g}_{D_i}(z_t, z_{t-1}) = \frac{1}{(n_{D_i}-1)h_1 h_2} \sum_{j=2}^{n_{D_i}} K_1\left(\frac{z_t - w_{ij}}{h_1}\right) K_2\left(\frac{z_{t-1} - w_{i,j-1}}{h_2}\right)$$

and

$$\hat{r}_{D_i}(z_{t-1}) = \frac{1}{(n_{D_i}-1)h_3} \sum_{j=2}^{n_{D_i}} K_3\left(\frac{z_{t-1} - w_{i,j-1}}{h_3}\right).$$

Here, h_1, h_2 and h_3 are the bandwidths, and K_1, K_2 and K_3 are the kernel functions. See Fan et al. (1996), Hall et al. (2004) and Silverman (1986) for further details. In this analysis, the Gaussian kernel

$$K(s) = (2\pi)^{-1/2} \exp(-s^2/2)$$

is used for all three functions K_1, K_2 and K_3.

The functions $\hat{f}_{D_i}$ for each $i \in \{1, 2, \ldots, m_D\}$ can be calculated in R using the np package (Hayfield and Racine (2008)). This package sets $h_2 = h_3$ (the bandwidths that apply to the previous observation in the numerator and denominator, respectively), and finds the optimal bandwidths h_1 and h_2 using leave-one-out cross-validation and maximising the estimated likelihood, a method which is described in Silverman (1986). The idea behind leave-one-out cross-validation (when maximising the likelihood) is that the conditional density function should be estimated, leaving out each pair of the observed data at a time. The estimated log-likelihood of the removed pair can then be calculated from the estimated conditional density function. The average of these estimated log-likelihoods can then be maximised with respect to the bandwidths h_1 and h_2. More formally, the function

$$CV(h_1, h_2) = \frac{1}{n_{D_i} - 1} \sum_{j=1}^{n_{D_i}-1} \log \hat{f}_{-j,D_i}(w_{ij} \mid w_{i,j-1})$$

can be maximised to find the values of h_1 and h_2 that maximise the estimated log-likelihood. The notation $\hat{f}_{-j,D_i}(\cdot \mid \cdot)$ indicates that the jth pair $(w_{ij}, w_{i,j-1})$ was not included in the estimation of the conditional density function $\hat{f}_{D_i}$. The maximising of the function $CV(h_1, h_2)$ is done numerically.

The bandwidths h_1 and h_2 are known as fixed bandwidths because they remain constant for all values of w_{ij} and $w_{i,j-1}$. When using fixed bandwidths, there are well documented (see for example p. 18 of Silverman (1986)) problems with using kernel density estimates to estimate the tails of density functions, where few data may be present. Better results may be obtained using a bandwidth which varies, depending on the amount of data nearby. Therefore, two different bandwidth types are considered for the calculation of the functions $\hat{f}_{D_i}$, for comparison. The first type is a fixed bandwidth, in which h_1, h_2 and h_3 remain constant at all values of w_{ij} and $w_{i,j-1}$. The second type is an adaptive

nearest neighbour bandwidth, introduced in Breiman et al. (1977). This type of bandwidth will vary, depending on the amount of data close by, becoming larger as the amount of nearby data reduces. Using an adaptive nearest neighbour bandwidth, the kernel density estimate, $\hat{r}_{D_i}(z_{t-1})$ becomes

$$\hat{r}_{D_i}(z_{t-1}) = \frac{1}{(n_{D_i}-1)} \sum_{j=2}^{n_{D_i}} \frac{1}{h_{3j}(k_3)} K_3\left(\frac{z_{t-1} - w_{i,j-1}}{h_{3j}(k_3)}\right)$$

where $h_{3j}(k_3)$ is the Euclidean distance from the point $w_{i,j-1}$ to the k_3-th nearest data point (see Terrell and Scott (1992) for further details). The kernel density estimate, $\hat{g}_{D_i}(z_t, z_{t-1})$, with bandwidths h_1 and h_2, changes similarly. Leave-one-out cross-validation is used to select the values of k_1 (associated with h_1) and k_2 (associated with h_2) that maximise the estimated log-likelihood (since, as before, $h_2 = h_3$).

Given $\hat{g}_{D_i}$ and $\hat{r}_{D_i}$, estimates of the conditional density function $\hat{f}_{D_i}(\cdot \mid \cdot)$ can be obtained using (3.18). This estimated conditional density function must be calculated for each of the samples in B and C for both a fixed and an adaptive bandwidth.

3.5 Conclusion

In this chapter, it was shown how to fit three different models, each accounting for autocorrelation at lag one, to a sample of autocorrelated data. The three models were an autoregressive process of lag one (both with and without random effects), a hidden Markov model and a nonparametric model, with both fixed and variable bandwidth. The autoregressive model and the nonparametric model account for lag one autocorrelation. In addition, the hidden Markov model allows the data to be driven by an underlying latent Markov chain. Using the hidden Markov model, autocorrelation arising from dependence between observations can be modelled alongside the correlation arising from the clustering of observations with similar mean and variance levels. By setting the autocorrelation parameter to zero in the hidden Markov model, it is possible to model just the clustering alone. Two of the models described (the autoregressive model and the hidden Markov model) are parametric, and make an assumption of Normality of the error terms. The nonparametric model dispenses with this assumption, and uses kernel density techniques to estimate conditional density functions.

It was described how to fit these three models to two training sets of data, B and C. The aim is to use parameter estimates from the parametric models, or function estimates from the nonparametric model, to evaluate the likelihood of questioned samples of data belonging to each of these two sets. In the context of evidence evaluation, the training set B relates to a proposition H_B, which could be that the questioned sample is from the background population (so that B would consist of m_B samples from the background population). The training set C relates to a different proposition H_C, which could be that the questioned sample is involved with crime (so that C would consist of m_C samples involved with crime). The aim is to evaluate the likelihood of both H_B and H_C, conditional on the evidence, or the questioned sample. The parameter estimates, which consist of draws from the

posterior distribution of the parameters, conditional on the training data, or function estimates, for each of these sets, B and C, are required to evaluate these two likelihoods. The methods described in this chapter allow these parameter and function estimates to be obtained. In the following chapter, these parameter and function estimates will be used to evaluate the likelihood of both H_B and H_C, conditional on the questioned sample, for each of the models described.

Chapter 4

Evaluating the likelihood ratio for autocorrelated data

Where evidence has been taken from a crime scene, part of that evidence may be a sample of autocorrelated data, known as the questioned sample. The measurements on this questioned sample, known as the evidential data, are given by $\mathbf{z} = (z_1, \ldots, z_n)$. The value of evidence, or likelihood ratio, associated with the propositions of association with crime (H_C) and of association with the background population (H_B) for this questioned sample is given by

$$\frac{f(\mathbf{z} \mid H_C)}{f(\mathbf{z} \mid H_B)}.$$

If this statistic is greater than one, then the evidence is said to support proposition H_C. If it is less than one, then the evidence is said to support proposition H_B. The logarithm of the likelihood ratio is known as the weight of evidence. The absolute value of the weight of evidence measures the extent of support given by the evidence to the proposition it supports.

If a parametric model is used to model $\mathbf{z}$, let the parameter θ^C (possibly multivariate) characterize the probability density function of samples known to be from training set C (associated with H_C) and let the parameter θ^B characterize the probability density function of samples known to be from training set B (associated with H_B). The parameters θ^C and θ^B can be estimated using their associated training sets. Conditioning on the two parameters θ^C and θ^B, the likelihood ratio becomes

$$\frac{f(\mathbf{z} \mid H_C)}{f(\mathbf{z} \mid H_B)} = \frac{\int f(\mathbf{z} \mid \theta^C) f(\theta^C \mid \mathbf{y}) \, \mathrm{d}\theta^C}{\int f(\mathbf{z} \mid \theta^B) f(\theta^B \mid \mathbf{x}) \, \mathrm{d}\theta^B}.$$

The functions $f(\theta^C \mid \mathbf{y})$ and $f(\theta^B \mid \mathbf{x})$ are known as between sample density functions, as they allow for differences in the value of θ^C and θ^B between different samples. The training data $\mathbf{y}$ and $\mathbf{x}$ are explicitly conditioned on in these functions to indicate which set of training data is used to obtain each between sample density function. Elsewhere, the training data $\mathbf{x}$ and $\mathbf{y}$ are assumed to be

part of the background information, and the dependence of the likelihood ratio on this background information is not explicitly noted.

In this chapter, methods for the evaluation of likelihood ratios for the three models presented in Chapter 3 for autocorrelated data, are described. These methods allow the likelihood ratio to be evaluated for three different types of autocorrelated evidential data. Firstly, data which can be modelled with an autoregressive process of order one (both with and without random effects). Secondly, data which can be modelled with a hidden Markov model that does not assume independence of observations, conditional on the hidden states. Lastly, data which can be modelled using a nonparametric model that allows for lag one autocorrelation. The autoregressive model and the hidden Markov model are two-level hierarchical models; they model variation in the mean, variance, autocorrelation and transition probabilities (in the case of the hidden Markov model) between samples. The nonparametric model accounts for differences in the probability density function between samples.

A fourth model, known as the standard model, assumes independence between observations within a sample. The method for the evaluation of the likelihood ratio for this model is described so that comparisons can be made between models with and without the independence assumption.

In the literature relating to forensic evidence interpretation, which is summarised in Section 2.1.2, the form of the likelihood ratio for the problem of comparing the source of two evidential samples has been given for two-level hierarchical models with univariate, independent and Normally distributed data with a Normally distributed between sample distribution on the mean (Lindley (1977)), an exponentially distributed between sample distribution on the mean (Aitken, Shen et al. (2007)) and with a kernel density estimate used for the between sample distribution on the mean (Aitken and Taroni (2004)). Alberink et al. (2013) extended this work to allow for a non-constant variance between samples. Likelihood ratios have also been calculated for independent data with a multivariate Normal distribution; variation in the mean between samples is modelled with either a multivariate Normal distribution or with a multivariate kernel density estimate (Aitken and Lucy (2004)). In Bozza et al. (2008), this work was extended by allowing for non-constant covariance between different samples. The problem considered in this thesis is the discrimination problem introduced in Section 2.1.5. The approaches described above for the comparison of sources problem can be adapted to evaluate the likelihood ratio for the discrimination problem, although the limitations and assumptions of these approaches will still apply.

It is possible to model an autoregressive process of lag one using a multivariate Normal distribution. The vector $\mathbf{z} = (z_1, \ldots, z_n)$, following an autoregressive process of lag one so that

$$z_t - \mu = \alpha\,(z_{t-1} - \mu) + \epsilon_t$$

where $t = 2, \ldots, n$, $\epsilon_t \sim N(0, \sigma^2)$, $z_1 \sim N(\mu, \sigma^2/(1-\alpha^2))$ and $|\alpha| < 1$, can be shown to have a multivariate Normal distribution with n-dimensional mean $\mu = (\mu, \mu, \ldots, \mu)$ and $n \times n$ dimensional covariance matrix

$$\frac{\sigma^2}{(1-\alpha^2)}\begin{pmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{(n-1)} \\ \alpha & 1 & \alpha & \dots & \alpha^{(n-2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha^{(n-1)} & \alpha^{(n-2)} & \alpha^{(n-3)} & \dots & 1 \end{pmatrix}.$$

Due to the constrained form of this covariance matrix, which depends on only two parameters, the methods used in Bozza et al. (2008) are not appropriate for the evaluation of the likelihood ratio for $\mathbf{z}$. This is because the between sample distribution for this covariance matrix cannot be modelled with an inverse Wishart distribution, as done in Bozza et al. (2008). By assuming a constant variance σ^2 and a constant autocorrelation parameter α between samples, methods similar to those in Aitken and Lucy (2004) can, however, be applied. In this chapter, methods for evaluating the likelihood ratio for this autoregressive process, without assuming a constant autocorrelation parameter α, variance σ^2 or mean μ between different samples are given. These methods are then extended to evaluate likelihood ratios for the hidden Markov model.

As discussed in Section 2.1.2, in Lindley (1977), Aitken and Lucy (2004), Aitken and Taroni (2004), Aitken, Shen et al. (2007) and Bozza et al. (2008), the parameters of the between sample distribution are estimated using summary statistics, obtained from the training data sets B and C. Alberink et al. (2013) use a Bayesian approach to obtain estimates of the parameters of the between sample distribution. Prior distributions are placed on the parameters of the between sample distribution, and these prior distributions are combined with the training data to obtain draws from the posterior distributions of the parameters. These draws are then used in the evaluation of the likelihood ratio. In this chapter, a Bayesian approach is used, as in Alberink et al. (2013). For the autoregressive model with random effects, the method is the same as that in Alberink et al. (2013), just with the addition of a parameter to model the autocorrelation. For the autoregressive model without random effects and the hidden Markov model, in Chapter 3, draws were obtained from the posterior distribution of the multivariate parameter θ_i, conditional on each individual sample i, for samples in both B and C (where θ_i is either equal to $\theta^D_{H_i}$ or $\theta^D_{A_i}$ for $D = B$ or $D = C$). In this chapter, these individual posterior distributions are combined to form an overall distribution for θ, based on the entire training data set. This is done by summing and weighting the individual posterior distributions; the method is described later in this chapter.

Likelihood ratios were evaluated using a nonparametric within sample distribution for the data $\mathbf{z}$ in Besson (2004). Examples of the evaluation of likelihood ratios using nonparametric methods can also be seen in Aitken and Taroni (2004) and Taroni, Aitken et al. (2006). This previous work has assumed independent data, so that standard kernel density estimates can be used to model the probability density function of the data under each of the two propositions. In this chapter, kernel density estimates of conditional probability density functions are used, so that autocorrelation can be taken into account.

As in Chapter 3, a general notation is used, for a general training data set D, containing data $\mathbf{w}$. The notation $\mathbf{w}$, without a subscript, is used to denote either the entire data set $\mathbf{x}$ or the entire data set $\mathbf{y}$. The notation $\mathbf{w}_i$ refers to a single sample from the data set $\mathbf{w}$. Methods are given for the evaluation of the likelihood of H_D, given by $f(\mathbf{z} \mid H_D)$. For the evaluation of the likelihood ratio, the ratio of the likelihood of H_C and the likelihood of H_B must be calculated.

4.1 The autoregressive model

Methods for obtaining draws from the posterior distributions of $\theta^C_{A_i}$ and $\theta^B_{A_i}$, which are the parameters for the autoregressive model, conditional on the i-th sample from $\mathbf{y}$ and $\mathbf{x}$ respectively, were given in Chapter 3. The associated parameters that will be used to evaluate the likelihood ratio for the autoregressive model for the questioned sample $\mathbf{z}$ are given by θ^C_A and θ^B_A.

Two aspects of the calulation of the likelihood ratio are described in this section. The first is the use of posterior density functions of the parameters of individual samples ($f(\theta^D_{A_i} \mid \mathbf{w}_i)$) to estimate the overall posterior density function $f(\theta^D_A \mid \mathbf{w})$ (also known as the between sample density function) , which is required to calculate the likelihood ratio. The second aspect is the calculation of the likelihood ratio using this between sample density function.

Conditioning on the parameter vector θ^D_A, the likelihood of proposition H_D can be written as

$$f(\mathbf{z} \mid H_D) = \int f(z_1, z_2, \ldots, z_n \mid \theta^D_A) f(\theta^D_A \mid \mathbf{w}) \, \mathrm{d}\theta^D_A.$$

This integral is approximated using a weighted sum of m_D likelihoods of proposition H_D, where each of these likelihoods is calculated using one of the m_D samples in training set D. This approximation relies on the m_D samples in D being both independent and a representative sample of the whole population. The likelihood $f(\mathbf{z} \mid H_D)$ is estimated by

$$f(\mathbf{z} \mid H_D) \approx \sum_{i=1}^{m_D} P(\theta^D_A = \theta^D_{A_i}) \int f(z_1, z_2, \ldots, z_n \mid \theta^D_{A_i}) f(\theta^D_{A_i} \mid \mathbf{w}_i) \, \mathrm{d}\theta^D_{A_i}.$$

The measurements $\mathbf{z}$ are assumed to follow an autoregressive process of order one. Therefore, conditional on the $(t-1)$-th observation, the t-th observation is independent of observations $1, 2, \ldots, t-2$ for $t \in \{2, \ldots, n\}$. As a result, the likelihood can be written

$$f(\mathbf{z} \mid H_D) \approx \sum_{i=1}^{m_D} v_i \int f(z_1 \mid \theta^D_{A_i}) f(z_2 \mid z_1, \theta^D_{A_i}) \ldots f(z_n \mid z_{n-1}, \theta^D_{A_i}) f(\theta^D_{A_i} \mid \mathbf{w}_i) \, \mathrm{d}\theta^D_{A_i}. \tag{4.1}$$

Here, the probability $P(\theta^D_A = \theta^D_{A_i})$ has been replaced by a weight v_i. A set of suggested values for these weights is given by $v_i = n_{D_i} / \sum_{i=1}^{m_D} n_{D_i}$. The use of these suggested weights results in samples with a larger number of observations having a greater influence on the likelihood. A discussion of the effect that this choice of weights has on the likelihood ratio for data relating to traces of cocaine on banknotes is given in Section 6.5.7.

As discussed in the introduction, previous work (Lindley (1977); Aitken and Lucy (2004); Aitken and Taroni (2004); Bozza et al. (2008); Aitken, Shen et al. (2007)), has determined the between sample distribution using summary statistics of the training data. The equivalent here would be calculating point estimates of the parameters making up $\theta^D_{A_i}$ and using either a kernel density estimate or a parametric density function based on these point estimates, to model the between sample density, $f(\theta^D_A \mid \mathbf{w})$. For example, in the case of the parameter $\mu^D_{A_i}$, the posterior mean of $\mu^D_{A_i}$, denoted $\bar{\mu}^D_{A_i}$, could be calculated for each sample. Then a Normal distribution could be used as the between sample distribution of μ^D_A, with mean taken to be $(1/m_D)\sum \bar{\mu}^D_{A_i}$ and variance taken to be $(1/(m_D-1))\sum(\bar{\mu}^D_{A_i} - (1/m_D)\sum \bar{\mu}^D_{A_i})^2$. The other parameters could be modelled similarly by, for example, using the distributions listed as prior distributions in Section 3.1.1 with the parameters estimated appropriately. The product of these distributions for each of the parameters would then be used as the between sample distribution. It is not possible, due to the variation of the autocorrelation parameter between samples, to obtain an analytical solution for the likelihood ratio for the autoregressive model using this method (as done for independent data in Aitken and Lucy (2004) and Lindley (1977)). Instead, a sampler similar to that used in Bozza et al. (2008), using the conditional densities set out in Appendix A.1, could be used to estimate the likelihood ratio.

One problem with the use of this method for the models described here is that the parameters in θ^D_A might not be independent given the data $\mathbf{w}$. When estimating the between sample distribution as described in the previous paragraph, either an assumption of independence of the parameters would have to be made, or a potentially complicated multivariate between sample distribution would have to be specified. Sampling from the posterior distribution of the multivariate parameter $\theta^D_{A_i}$, as done in Chapter 3 for both the Gibbs and the Metropolis-Hastings samplers, avoids making this independence assumption.

Considering (4.1), the between sample density function $f(\theta^D_A \mid \mathbf{w})$ can be written as

$$f(\theta^D_A \mid \mathbf{w}) = \sum_{i=1}^{m_D} v_i f_i(\theta^D_A \mid \mathbf{w}_i). \tag{4.2}$$

Here, the posterior density function of the parameter $\theta^D_{A_i}$, evaluated at the point θ^D_A, has been written as $f_i(\theta^D_A \mid \mathbf{w}_i)$, instead of $f(\theta^D_A \mid \mathbf{w}_i)$ (this latter form is consistent with the notation in (4.1)). This change in notation is so that the posterior density functions associated with different samples are clearly differentiated. The between sample density function in (4.2) is the density function associated with a finite mixture distribution; the probability of selection of each of the m_D probability density functions f_i is given by the weights v_i. Any sample with a large weight will have a large bearing on the overall distribution. By incorporating the entire posterior density function of each parameter $\theta^D_{A_i}$ into the between sample density function, instead of using a multivariate kernel density estimate of the posterior mean for each sample, the variance of $\theta^D_{A_i}$ is taken into account.

Draws from the posterior density functions $f(\theta^D_{A_i} \mid \mathbf{w}_i)$ for $i \in \{1,\ldots,N\}$ were obtained in Chapter 3. Denote the rth draw from $f(\theta^D_{A_i} \mid \mathbf{w}_i)$ by $\theta^{D(r)}_{A_i}$, where $r \in \{1,2,\ldots,N\}$. Each draw $\theta^{D(r)}_{A_i}$ consists of

the parameters $(\mu_i^{D(r)}, \alpha_i^{D(r)}, (\sigma_i^{D(r)})^2)$ (the parameter $\beta_i^{D(r)}$ is not required in the calculation of the likelihood ratio). Each integral in (4.1) can be estimated separately using Monte Carlo integration by

$$\begin{aligned}&\int f(z_1 \mid \theta_{A_i}^D) f(z_2 \mid z_1, \theta_{A_i}^D) \dots f(z_n \mid z_{n-1}, \theta_{A_i}^D) f(\theta_{A_i}^D \mid \mathbf{w}_i)\, \mathrm{d}\theta_{A_i}^D \\ &\approx \frac{1}{N} \sum_{r=1}^{N} f(z_1 \mid \theta_{A_i}^{D(r)}) f(z_2 \mid z_1, \theta_{A_i}^{D(r)}) \dots f(z_n \mid z_{n-1}, \theta_{A_i}^{D(r)}). \end{aligned} \tag{4.3}$$

The functions $f(z_1 \mid \theta_{A_i}^{D(r)})$ and $f(z_t \mid z_{t-1}, \theta_{A_i}^{D(r)})$ for $t \in \{2, \dots, n\}$ are given by the definition of the autoregressive process in (3.2). The function $f(z_1 \mid \theta_{A_i}^{D(r)})$ is the probability density function of the Normal distribution, with mean $\mu_i^{D(r)}$ and variance $(\sigma_i^{D(r)})^2$. The functions $f(z_t \mid z_{t-1}, \theta_{A_i}^{D(r)})$ are also given by the probability density function of the Normal distribution, this time with mean $\mu_i^{D(r)} + \alpha_i^{D(r)}(z_{t-1} - \mu_i^{D(r)})$ and variance $(\sigma_i^{D(r)})^2$.

Having obtained estimates for each of the integrals in (4.1), these estimates can be combined to estimate the overall likelihood $f(\mathbf{z} \mid H_D)$. This process should be repeated to estimate both $f(\mathbf{z} \mid H_C)$ and $f(\mathbf{z} \mid H_B)$. The ratio of these two likelihoods gives the likelihood ratio. The value obtained for the likelihood ratio is an estimate, so care should be taken with the choice of N, to ensure that the variance of the estimate is not too large. Denote

$$f^{(r)}(\mathbf{z} \mid H_D) = \sum_{i=1}^{m_D} v_i f(\mathbf{z} \mid \theta_{A_i}^{D(r)})$$

so that the Monte Carlo estimate of the likelihood $f(\mathbf{z} \mid H_D)$ is given by $(1/N)\sum_{r=1}^{N} f^{(r)}(\mathbf{z} \mid H_D)$. Then, an estimate of the variance of the values of $f^{(r)}(\mathbf{z} \mid H_D)$ is given by

$$V = \frac{1}{N-1} \sum_{r=1}^{N} \left[f^{(r)}(\mathbf{z} \mid H_D) - \frac{1}{N} \sum_{r=1}^{N} f^{(r)}(\mathbf{z} \mid H_D) \right]^2$$

and hence a 95% confidence interval around the estimate of the likelihood is given by

$$(1/N) \sum_{r=1}^{N} f^{(r)}(\mathbf{z} \mid H_D) \pm 1.96 \sqrt{\frac{V}{N}}.$$

This confidence interval assumes Normality of the N values of $f^{(r)}(\mathbf{z} \mid H_D)$. If there are a large number of observations (i.e. n is large), practical experience suggests that this assumption of Normality may be violated. A more robust 95% confidence interval can be obtained by estimating $f(\mathbf{z} \mid H_D)$ multiple times, ordering the estimates, and taking the 2.5% and 97.5% quantiles of the ordered estimates as the lower and upper confidence limits. This estimate of a confidence interval requires a large number of estimates of $f(\mathbf{z} \mid H_D)$ to be obtained. The computational burden of this could potentially be large. More information on the use of these estimates in practice is given in Chapter 6. More information on using Monte Carlo integration to estimate integrals can be seen in Robert and Casella (2004) and Gelman, Carlin et al. (2004).

There has been some discussion in the literature regarding the use of intervals for estimates of likelihood ratios in forensic science, rather than point estimates (e.g. see Alberink et al. (2013) and Morrison (2011)). It is necessary to consider intervals for the likelihood ratio estimates described here because Monte Carlo integration has been used to obtain the estimate of the likelihood ratio. In many previous applications in forensic science (e.g. Aitken and Lucy (2004)), the likelihood ratio can be calculated exactly. Using Monte Carlo integration adds additional errors into the calculation that have not been accounted for elsewhere.

4.2 The autoregressive model with random effects

In this section, the method for obtaining the likelihood ratio for the autoregressive model with random effects, described in Section 3.2, is given. In Section 3.2, draws were obtained from the posterior distributions of the parameters $\theta^B_{A_r} = (\mu^B_\mu, \sigma^B_\mu, \gamma^B_V, \beta^B_V, \mu^B_\alpha, \sigma^B_\alpha)$ and $\theta^C_{A_r} = (\mu^C_\mu, \sigma^C_\mu, \gamma^C_V, \beta^C_V, \mu^C_\alpha, \sigma^C_\alpha)$, conditional on the data sets $\mathbf{x}$ and $\mathbf{y}$, respectively. The approach used to evaluate the likelihood ratio in Section 4.1 was to obtain the overall posterior density function $f(\theta^D_A \mid \mathbf{w})$ by summing weighted individual posterior density functions $f(\theta^D_{A_i} \mid \mathbf{w}_i)$. For the autoregressive model with random effects, Markov chain Monte Carlo methods were described in Section 3.2 for obtaining draws from the overall posterior density function $f(\theta^D_{A_r} \mid \mathbf{w})$ from the outset. Therefore, the likelihood of proposition H_D can be written directly as

$$\begin{aligned} f(\mathbf{z} \mid H_D) &= \int\int\int\int f(\mathbf{z} \mid \mu_z, \alpha_z, \sigma_z^2) f(\mu_z, \alpha_z, \sigma_z^2, \theta^D_{A_r} \mid \mathbf{w}) \, \mathrm{d}\theta^D_{A_r} \mathrm{d}\mu_z \mathrm{d}\alpha_z \mathrm{d}\sigma_z^2 \\ &= \int\int\int\int f(\mathbf{z} \mid \mu_z, \alpha_z, \sigma_z^2) f(\mu_z, \alpha_z, \sigma_z^2 \mid \theta^D_{A_r}) f(\theta^D_{A_r} \mid \mathbf{w}) \, \mathrm{d}\theta^D_{A_r} \mathrm{d}\mu_z \mathrm{d}\alpha_z \mathrm{d}\sigma_z^2 \end{aligned} \tag{4.4}$$

where μ_z, σ_z^2 and α_z are the mean, variance and autocorrelation parameter for the questioned sample $\mathbf{z}$, with probability distributions defined as for μ_i, σ_i^2 and α_i in (3.7), (3.8) and (3.9). To estimate this integral, the distributions given in (3.7), (3.8) and (3.9) can be used to obtain draws $\mu_z^{(r)}$, $\sigma_z^{2(r)}$ and $\alpha_z^{(r)}$, conditional on the r-th draw from $\theta^D_{A_r}$, given by $\theta^{D(r)}_{A_r}$. Using Monte Carlo integration, the likelihood of H_D is then estimated by

$$f(\mathbf{z} \mid H_D) \approx \frac{1}{N} \sum_{r=1}^{N} f(\mathbf{z} \mid \mu_z^{(r)}, \alpha_z^{(r)}, \sigma_z^{2(r)}),$$

where N is the total number of draws available from the density $f(\theta^D_{A_r} \mid \mathbf{w})$.

This method of estimating the likelihood ratio seems theoretically more satisfying than the method given in Section 4.1, which requires the estimation of the between sample density function from a weighted sum of the posterior density functions for each individual sample. There are, however, several practical disadvantages. Firstly, there are more parameters to sample from, which can make the implementation of a sampler more difficult. This did not cause difficulties with the autoregressive

model for data relating to traces of cocaine on banknotes, but for more complicated models such as the hidden Markov model it could be problematic. Also, the process of obtaining draws from the posterior distribution of the parameters uses the entire set of training data **w**. If this training data set is large, then there will be limitations caused by the amount of computing power available. Further computational problems arise because any sampler used to obtain draws from the posterior distribution of the parameters $\theta^D_{A_r}$ must be re-run every time another sample is added to the training set D. The method of summing weighted individual posterior distributions only requires the sampler to be run for the sample that is being added to the data set. A disadvantage relating to the model fit is that the distributions of μ_i, σ_i^2 and α_i must be parametric distributions. For data relating to traces of cocaine on banknotes, the distribution of μ_i was not thought to be Normal. The method of summing weighted individual posterior distributions, used in Sections 4.1, 4.3 and 4.4, has similarities to nonparametric methods using kernel density estimates; as such, this method is not reliant on the specification of a parametric between sample distribution. Lastly, the dimension of the integral in (4.4) which relates to the autoregressive process with random effects, is larger than the dimension of the integrals in (4.1) which relate to the autoregressive process without random effects. Estimation methods are required to evaluate these integrals, and so the increased dimension may lead to a decline in accuracy. Because of these practical disadvantages, the random effects model for the hidden Markov model, which has more parameters than that of the autoregressive model, was not developed.

4.3 The hidden Markov model

In this section, the methods given in Section 4.1 are extended for the evaluation of the likelihood of H_D, using the hidden Markov model. The arguments used to obtain (4.1) can also be used when the hidden Markov model is under consideration, but replacing θ^D_A and $\theta^D_{A_i}$ by θ^D_H and $\theta^D_{H_i}$ respectively, where the subscript H indicates that the parameters are those of the hidden Markov model. The approximation of the likelihood of H_D for the hidden Markov model is therefore given by

$$f(\mathbf{z} \mid H_D) \approx \sum_{i=1}^{m_D} \upsilon_i \int f(\mathbf{z} \mid \theta^D_{H_i}) f(\theta^D_{H_i} \mid \mathbf{w}_i) \, \mathrm{d}\theta^D_{H_i}. \tag{4.5}$$

As in Section 4.1, weights υ_i are used to represent the probability $P(\theta^D_H = \theta^D_{H_i})$, with a set of suggested weights being given by $\upsilon_i = n_{D_i} / \sum_{i=1}^{m_D} n_{D_i}$. As before, the between sample distribution is equivalent to a finite mixture distribution.

As in (4.3), each of the integrals in (4.5) can be approximated used Monte Carlo integration. Using the same notation as in Section 4.1, with $\theta^{D(r)}_{H_i}$ for $r \in \{1, \ldots, N\}$ representing the r-th draw from the posterior density function $f(\theta^D_{H_i} \mid \mathbf{w}_i)$ (methods for obtaining these draws are given in Section 3.3.6), each integral in (4.5) can be approximated by

$$\int f(\mathbf{z} \mid \theta_{H_i}^D) f(\theta_{H_i}^D \mid \mathbf{w}_i)\, \mathrm{d}\theta_{H_i}^D$$
$$\approx \frac{1}{N} \sum_{r=1}^{N} f(\mathbf{z} \mid \theta_{H_i}^{D(r)}). \tag{4.6}$$

In the case of the hidden Markov model, it is not straightforward to calculate the likelihood $f(\mathbf{z} \mid \theta_{H_i}^{D(r)})$, due to the added complication of the hidden states. A method for calculating this likelihood was given in Section 3.3.5. To use this method, $\mathbf{w}_i$ should be replaced with $\mathbf{z}$ and θ by $\theta_{H_i}^{D(r)}$.

As with the autoregressive model, the variance of the estimate of $f(\mathbf{z} \mid H_D)$ should be monitored. In the example discussed in Chapter 6, this variance was often larger for the hidden Markov model than for the autoregressive model.

4.4 Using a combination of the autoregressive model and the hidden Markov model

The hidden Markov model defined in (3.13) can be thought of as an extension to the autoregressive model defined in (3.2). The autoregressive model is equivalent to the hidden Markov model with just one state. It may be the case that some of the samples in the sets of training data are best modelled with an autoregressive process, and some are best modelled with a hidden Markov model. In other words, the samples in the training data sets should be modelled with a hidden Markov model, but with the number of states varying between samples (either one state or four states, the latter corresponding to two different mean and variance levels). The choice of model can be thought of as an extra parameter, with two possible values, 'autoregressive model' or 'hidden Markov model'. Assuming that the samples in the training data are a random selection from the overall population, then this parameter can be estimated from the data. In this section, a method for the evaluation of the likelihood ratio in this situation is discussed.

Let $\mathbf{M} = (M_1, \ldots, M_{m_D})$ be the vector representing the model choice for the m_D samples in the general training data set D. Let $M_i = M_H$ if the model choice for the i-th sample is the hidden Markov model, and let $M_i = M_A$ if the model choice for the i-th sample is the autoregressive model of order one (without random effects). By comparing the marginal likelihoods of the models M_H and M_A, given by $f(\mathbf{w}_i \mid M_i = M_H)$ and $f(\mathbf{w}_i \mid M_i = M_A)$, the model which best fits the data $\mathbf{w}_i$ can be selected. Two methods for obtaining these marginal likelihoods are described in Appendix B. One is the method developed in Chib and Jeliazkov (2001), which relies on the draws obtained from the posterior distribution of the parameter θ, conditional on the data $\mathbf{w}_i$, obtained using the samplers discussed in Chapter 3. In order to use this method, the chain of draws from the sampler must have converged to the posterior distribution. In cases where this chain has not converged, a more straightforward Monte Carlo integration technique can be used to estimate the marginal likelihood. Using Monte

Carlo integration is computationally more time consuming, but removes the need to sample from the posterior distribution of the parameters for the model with the smaller marginal likelihood. For data relating to traces of cocaine on banknotes, convergence of the posterior distribution of $\theta^D_{H_i}$ conditional on a sample $\mathbf{w}_i$ was difficult to achieve when the model was overfitted for that sample (see Section 3.3.4 for details). By calculating the marginal likelihoods of the two models using Monte Carlo integration, the better fitting model for the i-th sample can be determined, without reliance on the draws from the posterior distribution of $\theta^D_{H_i}$. If the autoregressive model is found to be the better fitting model for this sample, then there is no need to obtain draws from the posterior density $f(\theta^D_{H_i} \mid \mathbf{w}_i)$ at all.

Given the marginal likelihoods of the two models, $f(\mathbf{w}_i \mid M_i = M_H)$ and $f(\mathbf{w}_i \mid M_i = M_A)$ for $i \in \{1, \ldots, m_D\}$, the model with the larger marginal likelihood can be chosen for each sample. If $M_i = M_H$ then the parameters used in the model for the i-th sample are $\theta^D_{H_i}$, which is the nine-dimensional vector of hidden Markov model parameters and if $M_i = M_A$ then the parameters used for the i-th sample are $\theta^D_{A_i}$, which is the four-dimensional vector of autoregressive model parameters. In the remainder of this section, the estimation of the likelihood ratio when there are two possible models for the questioned sample $\mathbf{z}$ is described.

Denote the random variable representing the model choice for the questioned sample $\mathbf{z}$ by M_z. Conditioning on M_z, the likelihood $f(\mathbf{z} \mid H_D)$ is given by

$$f(\mathbf{z} \mid H_D) = P(M_z = M_H \mid H_D) f(\mathbf{z} \mid H_D, M_z = M_H) + P(M_z = M_A \mid H_D) f(\mathbf{z} \mid H_D, M_z = M_A).$$

By conditioning on the model choice in this way, some of the model uncertainty is taken into account. The probabilities $P(M_z = M_H \mid H_D)$ and $P(M_z = M_A \mid H_D)$ depend on proposition H_D. If these probabilities are very different when different propositions are considered then the fit of each of the models to $\mathbf{z}$, conditional on the proposition, may help to discriminate between the two propositions. If instead, only the model which fits the questioned sample better is used, then this source of discrimination is lost.

Conditioning on the model parameters θ^D_H and θ^D_A, the likelihood can then be written

$$\begin{aligned} f(\mathbf{z} \mid H_D) =& P(M_z = M_H \mid H_D) \int f(\mathbf{z} \mid \theta^D_H, M_z = M_H) f(\theta^D_H \mid \mathbf{w}) \, \mathrm{d}\theta^D_H \\ &+ P(M_z = M_A \mid H_D) \int f(\mathbf{z} \mid \theta^D_A, M_z = M_A) f(\theta^D_A \mid \mathbf{w}) \, \mathrm{d}\theta^D_A. \end{aligned}$$

Using the same reasoning as in Section 4.1, this expression can be approximated by conditioning on the parameter for the questioned sample being equal to the parameters associated with each of the samples in the training data set, so that

$$f(\mathbf{z} \mid H_D) \approx P(M_z = M_H \mid H_D) \sum_{\substack{i=1 \\ i:M_i=M_H}}^{m_D} P(\theta_H^D = \theta_{H_i}^D) \int f(\mathbf{z} \mid \theta_{H_i}^D, M_z = M_H) f(\theta_{H_i}^D \mid \mathbf{w}_i)\, \mathrm{d}\theta_{H_i}^D$$
$$+ P(M_z = M_A \mid H_D) \sum_{\substack{i=1 \\ i:M_i=M_A}}^{m_D} P(\theta_A^D = \theta_{A_i}^D) \int f(\mathbf{z} \mid \theta_{A_i}^D, M_z = M_A) f(\theta_{A_i}^D \mid \mathbf{w}_i)\, \mathrm{d}\theta_{A_i}^D.$$

Finally, as before, the probabilities $P(M_z = M_H \mid H_D)P(\theta_H^D = \theta_{H_i}^D)$ and $P(M_z = M_A \mid H_D)P(\theta_A^D = \theta_{A_i}^D)$ are replaced by weights v_i, giving

$$f(\mathbf{z} \mid H_D) \approx \sum_{\substack{i=1 \\ i:M_i=M_H}}^{m_D} v_i \int f(\mathbf{z} \mid \theta_{H_i}^D, M_z = M_H) f(\theta_{H_i}^D \mid \mathbf{w}_i)\, \mathrm{d}\theta_{H_i}^D$$
$$+ \sum_{\substack{i=1 \\ i:M_i=M_A}}^{m_D} v_i \int f(\mathbf{z} \mid \theta_{A_i}^D, M_z = M_A) f(\theta_{A_i}^D \mid \mathbf{w}_i)\, \mathrm{d}\theta_{A_i}^D. \tag{4.7}$$

As before, a set of possible values for the weights are given by $v_i = n_{D_i} / \sum_{i=1}^{m_D} n_{D_i}$ so that the weight varies with the number of observations in the sample. With these weights, $P(M_z = M_H \mid H_D)$ is being approximated by the proportion of observations (of all observations associated with H_D) that are in samples being modelled with the hidden Markov model, e.g.

$$P(M_z = M_H \mid H_D) \approx \frac{\sum_{\substack{i=1 \\ i:M_i=M_H}}^{m_D} n_{D_i}}{\sum_{i=1}^{m_D} n_{D_i}},$$

with a similar approximation for $P(M_z = M_A \mid H_D)$. The probability $P(\theta_H^D = \theta_{H_i}^D)$ (when $M_i = M_H$) is then approximated by the proportion of observations (of all observations that are both associated with H_D and in samples that are being modelled using a hidden Markov model) that are in the i-th sample, e.g.

$$P(\theta_H^D = \theta_{H_i}^D) \approx \frac{n_{D_i}}{\sum_{\substack{i=1 \\ i:M_i=M_H}}^{m_D} n_{D_i}}.$$

The probability $P(\theta_A^D = \theta_{A_i}^D)$ has a similar approximation.

Using (4.7), the likelihood $f(\mathbf{z} \mid H_D)$ can be estimated using Monte Carlo integration with the draws from $f(\theta_{A_i}^D \mid \mathbf{w}_i)$ and $f(\theta_{H_i}^D \mid \mathbf{w}_i)$, as described in Sections 4.3 for the first integral and 4.1 for the second integral.

The method discussed here for evaluating the likelihood ratio can be used in other situations where the choice of model is uncertain, it is not restricted to the two models discussed. This method

can also be easily extended to cover situations with more than two possible models.

Other methods for the fitting of hidden Markov models with a variable number of states exist. In particular, the reversible jump algorithm (Green (1995); Green (2003)) is a Markov chain Monte Carlo algorithm that allows for jumps between different models which have different parameter dimensions. Using this algorithm a posterior distribution can be obtained for the number of hidden states. In this example, only two different models are considered, so the more simplistic approach of considering the marginal likelihood of the two models is feasible.

4.5 The nonparametric model

In this section, a method for the evaluation of the likelihood ratio for the nonparametric model given in Section 3.4 is described. Allowing for lag one autocorrelation, it is assumed that the probability density function for $\mathbf{z}$ (and likelihood of H_D) is given by

$$f(z_1, z_2, \ldots, z_n \mid H_D) = f(z_1 \mid H_D) f(z_2 \mid z_1, H_D) \ldots f(z_n \mid z_{n-1}, H_D). \tag{4.8}$$

There are m_D samples in the data set D. Assuming that the joint probability density function of each of these samples can also be written in the form seen in (4.8), then each sample has a conditional probability density function, $f_{D_i}(\cdot \mid \cdot)$ for $i \in \{1, \ldots, m_D\}$, and a marginal probability density function for the first observation, $f_{D_i}(\cdot)$, associated with it. Let v_i for $i \in \{1, \ldots, m_D\}$ define a set of weights, with $\sum v_i = 1$, such that v_i can be thought of as the probability that the adjacent pairs of observations in a random sample drawn from the population being considered have conditional density function $f_{D_i}(\cdot \mid \cdot)$ and that the first observation in the sample has marginal probability density function $f_{D_i}(\cdot)$. Then, using these weights, the likelihood of H_D can be approximated by

$$f(z_1, z_2, \ldots, z_n \mid H_D) \approx \sum_{i=1}^{m_D} v_i f_{D_i}(z_1 \mid H_D) f_{D_i}(z_2 \mid z_1, H_D) \ldots f_{D_i}(z_n \mid z_{n-1}, H_D). \tag{4.9}$$

In order to use this approximation, the m_D samples should be a representative sample of the population being considered. As for the parametric models, the weights could be defined by

$$v_i = \frac{n_{D_i}}{\sum_{i=1}^{m_D} n_{D_i}}$$

so that the weight for each sample increases as the number of observations in the sample increases.

A nonparametric method of estimating the functions f_{D_i} by $\hat{f}_{D_i}$ was described in Section 3.4. Substituting these estimates into (4.9) gives the approximation

$$\hat{f}(z_1, z_2, \ldots, z_n \mid H_D) \approx \sum_{i=1}^{m_D} v_i \hat{f}_{D_i}(z_1 \mid H_D) \hat{f}_{D_i}(z_2 \mid z_1, H_D) \ldots \hat{f}_{D_i}(z_n \mid z_{n-1}, H_D).$$

This approximation is used to estimate the likelihood of the proposition H_D using the questioned

sample **z**.

4.6 The standard model

The method for the evaluation of the likelihood ratio for a standard model, which assumes that observations within a sample are independent, is also described so that comparisons can be made between models with and without the independence assumption. This model is similar to that described in Aitken and Taroni (2004), from p. 330. A Normal distribution is assumed for the within sample distribution. The mean of this Normal distribution is allowed to vary between samples; a kernel density estimate is used to estimate this between sample distribution. The use of a kernel density estimate allows for non-Normal variation in the mean between samples. A slight adaptation of the method presented in Aitken and Taroni (2004) is required, because the problem being considered here is a discrimination problem, not a comparison problem (comparing of the source of a control and recovered item). The estimate of the between sample variance also has a slight adjustment, to allow for situations where the number of observations in each sample varies between samples. The likelihood ratio for a questioned sample **z**, with n banknotes, for the discrimination problem, is given by

$$\mathrm{LR} = \frac{\sum_{i=1}^{m_C}\left(m_C\sqrt{\tau_C^2+n\lambda_C^2 s_C^2}\right)^{-1}\exp\left[-\frac{n(\bar{\mathbf{z}}-\bar{\mathbf{y}}_i)^2}{2(\tau_C^2+n\lambda_C^2 s_C^2)}\right]}{\sum_{i=1}^{m_B}\left(m_B\sqrt{\tau_B^2+n\lambda_B^2 s_B^2}\right)^{-1}\exp\left[-\frac{n(\bar{\mathbf{z}}-\bar{\mathbf{x}}_i)^2}{2(\tau_B^2+n\lambda_B^2 s_B^2)}\right]} \tag{4.10}$$

where τ_C^2 and s_C^2 are respectively the within and between sample variances for samples in training set C and τ_B^2 and s_B^2 are the within and between sample variances for samples in training set B. The bandwidths of the kernel density estimates for the between sample distributions of the mean are given by $\lambda_C s_C$ (set C) and $\lambda_B s_B$ (set B). The within sample variance for samples in set C is estimated by

$$\hat{\tau}_C^2 = \sum_{i=1}^{m_C}\sum_{t=1}^{n_{C_i}}\frac{\left(y_{it}-\bar{\mathbf{y}}_i\right)^2}{N_C-m_C}$$

where

$$\bar{\mathbf{y}}_i = \sum_{t=1}^{n_{C_i}} y_{it}$$

and N_C denotes the total number of observations so that

$$N_C = \sum_{i=1}^{m_C} n_{C_i}.$$

The between sample variance for samples in set C is estimated by

$$\hat{s}_C^2 = \sum_{i=1}^{m_C} \frac{n_{C_i}\left(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}\right)^2}{\tilde{n}_C(m_C-1)} - \frac{\hat{t}_C^2}{\tilde{n}_C}$$

where the value of $\tilde{n}_C$ is given by

$$\tilde{n}_C = \frac{1}{m_C - 1}\left(N_C - \frac{\sum_{i=1}^{m_C} n_{C_i}^2}{N_C}\right).$$

The estimators used are the ANOVA estimators given on p. 19-21 of Snijders and Bosker (2012).

The bandwidth parameter λ_C is selected using Silverman's rule of thumb (Silverman (1986)), so that

$$\lambda_C = \left(\frac{4}{3m_C}\right)^{\frac{1}{5}}$$

The between and within sample variances and the parameter λ_B for samples in set B are estimated similarly by replacing y with x and C with B.

4.7 Conclusion

In this section, methods were given for the estimation of the likelihood ratio given by

$$\frac{f(\mathbf{z} \mid H_C)}{f(\mathbf{z} \mid H_B)}$$

where $\mathbf{z}$ is a sample of autocorrelated data of unknown source, and H_C and H_B are two competing propositions as to the origin of the data. The likelihood ratio was estimated for the three models discussed in Chapter 3: the autoregressive model of order one, the hidden Markov model and the nonparametric model. For the autoregressive model, the likelihood ratio was evaluated for models both with and without random effects. An adaptation of the method given in Aitken and Taroni (2004), for the evaluation of likelihood ratios for independent data, but for the discrimination problem, and allowing for a different number of observations in each sample, was also given. As such, models both with and without the independence assumption can be compared. The methods detailed in this chapter extend the work of Lindley (1977), Aitken and Lucy (2004), Bozza et al. (2008) and Alberink et al. (2013) to allow likelihood ratios to be calculated for data that are autocorrelated, or that are driven by an underlying and unobserved Markov chain. Previous methods have either assumed independence between observations or have required a full covariance matrix to be specified. Methods were also given in this chapter for the evaluation of the likelihood ratio when there is model uncertainty. Specifically, the likelihood ratio was estimated for the case when the questioned sample $\mathbf{z}$ might have been generated by either an autoregressive model or a hidden Markov model.

The method for evaluating the likelihood ratio for the autoregressive model (without random effects) and the hidden Markov model formed the between sample density function $f(\theta \mid \mathbf{w})$ from a weighted sum of the posterior density functions of θ, conditional on single samples $\mathbf{w}_i$. Draws from

these posterior distributions can be obtained using an MCMC sampler. The between sample density function is therefore specified by

$$f(\theta \mid \mathbf{w}) \approx \sum_{i=1}^{m_D} v_i f_i(\theta \mid \mathbf{w}_i),$$

for a set of weights v_i, where f_i is the posterior density function of θ, conditional on a single sample of training data $\mathbf{w}_i$. This method has both advantages and disadvantages when compared to the method using summary statistics described in Section 4.1. One advantage is that it allows subjective priors on the parameter θ to be used, so that expert opinion can be taken into account alongside the data. Another advantage is that it does not require summary statistics to be calculated, or the parametric form of the between sample distribution to be specified. Specifying the between sample distribution of a mean can be relatively straightforward, but the specification of a multivariate between sample distribution for the hidden Markov model parameters is more difficult. The method described here also does not assume independence of the individual parameters which comprise θ. The main disadvantage of the method using weighted sums of individual posterior distributions is the computational complexity, and the need for Markov chain Monte Carlo methods. However, since it is impossible to obtain an analytical solution for the likelihood ratio for all of the parametric models developed here which account for autocorrelation, computational methods would be required to obtain an estimate of the likelihood ratio in any case (as in Bozza et al. (2008)).

The method used to evaluate the likelihood ratio for the autoregressive model with random effects is similar to the method used in Alberink et al. (2013). This method has advantages when compared to the method using weighted sums of posterior distributions because draws from the between sample density function $f(\theta \mid \mathbf{w})$ can be obtained directly, without needing to approximate the between sample density with a weighted sum. However, there are also disadvantages, because the entire training data set must be used in one MCMC sampler, which for complicated models and large data sets can create computational problems. In addition, every time another sample is added to the training data set, the sampler used to obtain draws from $f(\theta \mid \mathbf{w})$ must be re-run, and the dimension of the integral that must be estimated to evaluate the likelihood ratio is larger, which can create problems with the accuracy of this estimate. Finally, using the method with weighted sums of posterior distributions avoids making a parametric assumption for the between sample distribution.

In the following chapters, the methods for the evaluation of likelihood ratios for autocorrelated data, developed in Chapters 3 and 4, will be applied to data relating to traces of cocaine on banknotes.

Chapter 5

Measurements of cocaine traces on banknotes as evidential data

5.1 Introduction

The models and methods given in Chapters 3 and 4 for the evaluation of likelihood ratios for autocorrelated data were motivated by evidential data relating to traces of drugs on banknotes. Banknotes can be seized from crime scenes (or from suspects) as evidence of the suspected involvement of a suspect with drug crime. Drugs can transfer from surfaces onto banknotes (Ebejer, Winn et al. (2007); Sleeman et al. (2000)) and so large quantities of a particular drug on a sample of banknotes may suggest that these banknotes have been in the vicinity of surfaces which are contaminated with that drug. In the absence of an illegal drug itself being found in the personal possession of a suspect, then drug traces on banknotes that belong to this suspect could provide evidence of a link between the suspect and activity involving the illegal drug. In Chapters 5 and 6, the connection of a suspect to a crime specifically involving the drug cocaine, where a sample of banknotes has been seized as evidence for this crime, is considered via the two propositions:

- H_C: that the banknotes are associated with a person who is involved with crime involving cocaine, and
- H_B: that the banknotes are associated with a person who is not involved with crime involving cocaine.

A measure of the quantity of cocaine on each banknote within a sample can be obtained. Samples of banknotes are generally found in bundles, and measurements can be obtained sequentially, so that the order of the measurements reflects the order of the banknotes within each bundle. Studies have been done (Ebejer, Winn et al. (2007)) which suggest that heroin can transfer from one banknote to another, suggesting that models which allow for autocorrelation might be appropriate for these data.

In this chapter, the data are introduced. The method used to obtain the raw data, which relate to the amount of drug measured by a mass spectrometer at a particular time, is explained and an algorithm developed to convert these raw data into a measure of the amount of drug on each banknote is described. The data validation and verification steps taken are outlined, and an analysis of the traits of the data is presented.

A key part of this chapter is Section 5.3, which concerns the selection of the training data sets and the formation of the propositions H_B and H_C, because the definitions used here differ from those used in previous treatments of similar data. The training set associated with H_C is formed of samples of banknotes that are associated with a person that has been convicted of a crime involving cocaine. The training set associated with H_B is formed of samples taken from general circulation. In Section 5.3, propositions H_B and H_C are defined more precisely than given above, and some of the limitations associated with this choice of propositions are discussed. Problems with the training datasets arise because banknotes that are associated with proposition H_C might not be contaminated any more than those associated with H_B, despite their association with a person who has been convicted of a crime involving cocaine. These problems, and the associated implications on the assessment of the models described in Chapter 3, are discussed.

5.2 Obtaining drug quantity measurements from banknotes

5.2.1 Tandem mass spectrometry

The amount of drug on a banknote is measured using a thermal desorption unit, fixed to a tandem mass spectrometer. This procedure is described in Ebejer, Brereton et al. (2005), Roberts et al. (1997) and Dixon et al. (2006) and the thermal desorption unit is shown in Figure 5.1. The thermal desorption unit consists of two heated plates (seen on the right hand side of the diagram). Analysts pass banknotes between these two heated plates one by one. Substances on the banknote evaporate, ionize and pass into the mass spectrometer. As described in Dixon et al. (2006), precursor ions are then selected by the mass spectrometer. These precursor ions correspond to the drugs being considered. For example the precursor ion of cocaine has mass to charge ratio *m/z 304*. These precursor ions are then fragmented. Two of the resulting fragments from each precursor ion (known as product ions) pass on to the detector. The transitions from the precursor ion to the product ion are known as gas phase ion transitions. The mass spectrometer measures the number of gas phase ion transitions detected and then converts this to a gas phase ion transition count per second. The gas phase ion transitions monitored for cocaine are $304 \rightarrow 105$ and $304 \rightarrow 182$. For the data considered here, the detector was generally set up to monitor ten different ion transitions, corresponding to two product ions from five different drugs, although for some of the samples used the detector was set up to monitor eight ion transitions, corresponding to four different drugs. The five drugs are cocaine, diamorphine (heroin), MDMA, amphetamine and THC (cannabis). Different ion transitions are monitored one after another;

one sweep through all ten or eight ion transitions is known as a scan. Ten such scans are carried out per second. Two different ion transitions are used for each drug because the presence of both ion transitions (with ratio of ion transition counts roughly as expected for the drug being monitored) allows the substance to be identified as the drug in question.

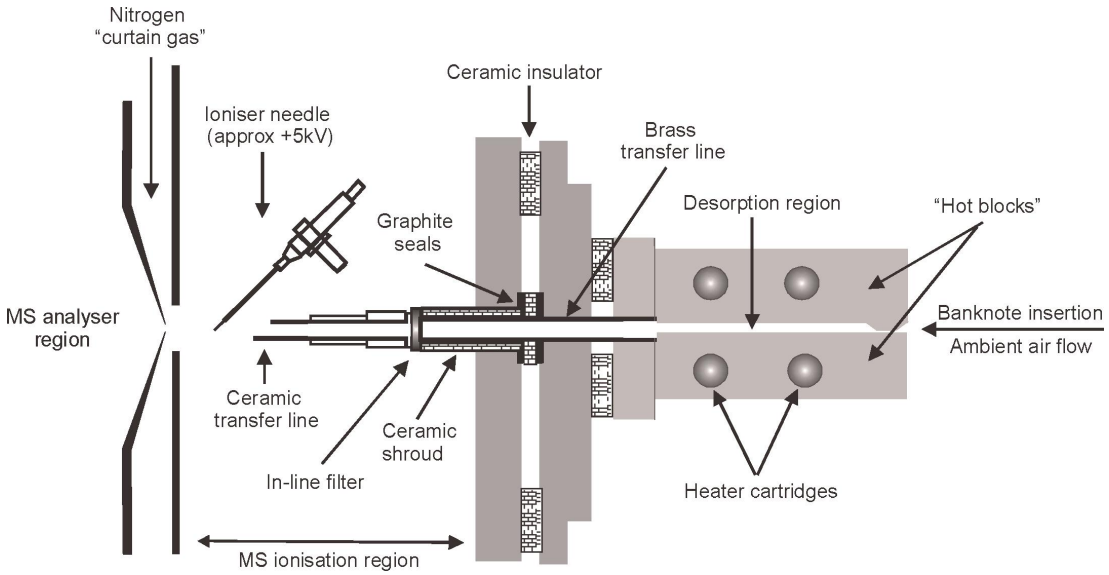

Figure 5.1: Diagram of thermal desorption unit used for analysis of banknotes. Reproduced with permission from Ebejer, Brereton et al. (2005).

The resulting output from the mass spectrometer gives a measure of the amount of each of the drugs on the banknotes tested, via the number of gas phase ion transitions detected. This measure is not an estimate of the exact amount of each drug on the banknotes. As stated in Armenta and Guardia (2008), the method of thermal desorption combined with tandem mass spectrometry is considered to be semi-quantitative because no method currently exists for extrapolating from the number of ion transitions detected to the amount of drug on the banknote. Despite this, it is still possible to use the data obtained in a statistical analysis because comparisons can be made between the number of ion transitions detected on different samples of banknotes.

The data used in this thesis arose from samples of banknotes analysed between the years 2002 and 2010. Each sample was analysed on one of four different machines. The choice of which machine to use in a particular situation was made for operational reasons. Regular tests were carried out on these machines to check that the response of each machine to a standard injection remained roughly constant over time. However, responses do vary slightly between machines. Attempts were made to account for these differences (see Section 5.5). In each of the two training data sets there are samples which were analysed on all four machines, and these analyses were done at a variety of different times, and by a variety of different people.

5.2.2 Other methods for measuring the quantity of drugs on banknotes

Other methods for measuring the quantity of drug on a banknote have been developed. In Jenkins (2001), GC-MS (gas chromatographic-mass spectrometry), which requires the drug first to be extracted from the banknote using a solvent, was used to test samples of US dollars from five cities. In Besson (2004) and Jourdan et al. (2013), IMS (ion mobility spectrometry), which involves either the banknotes being swabbed, or particles being collected onto a filter, was used to analyse Swiss banknotes and US banknotes respectively. Armenta and Guardia (2008) and Sleeman et al. (2000) both give a summary of the different analytical methods used to detect cocaine traces on banknotes. Conclusions are that the advantage of using a thermal desorption unit is that samples do not need to be prepared (e.g. by extracting the drug using a solvent) and so thermal desorption allows for much faster measurements of quantities of drugs on banknotes, meaning that whole samples can be tested. GC-MS is a more time consuming technique, and so it is only cost effective to analyse a small proportion of the banknotes in a sample. IMS is more rapid than GC-MS but IMS still requires the preparation of a sample, unlike thermal desorption. The disadvantage of thermal desorption in comparison to GC-MS is that thermal desorption is a so-called semi-quantitative method, so measurements taken are not a precise indication of the amount of drug on the banknote (as described in Section 5.2.1).

5.2.3 The banknote testing process

Samples of banknotes are either obtained from law enforcement agencies, having been brought in for testing in the course of a criminal case, or they are collected and tested to form part of a database of banknotes from general circulation. The samples, as often in forensic science, form a convenience sample (see p179 in Aitken and Taroni (2004), and p34 of Aitken, Roberts et al. (2010) for a discussion). The number of banknotes in any sample can vary. In the case of very large samples brought in by law enforcement agencies, not all of the banknotes are always analysed. The process for deciding which banknotes are analysed is partly subjective; it can depend on the wishes of the law enforcement agency, or the knowledge of the forensic expert analysing the sample.

Samples from law enforcement agencies can take many forms, but often consist of multiple bundles of banknotes which are secured with rubber bands, folded banknotes or bank strips. One sample of banknotes can contain multiple bundles. Often a bundle of banknotes is formed of multiple smaller sub-bundles. Selecting which of the banknotes will be analysed in large samples can depend on the form of the sample. In general, analysts will either try to test a number of banknotes from each bundle, selecting a number of notes from the top, middle and bottom of each bundle or they will select a number of smaller sub-bundles from each larger bundle and analyse each of these sub-bundles in their entirety. For example, a common situation is that the banknotes are in so-called 'dealer's wraps' (usually consisting of bundles of £100 with the outer note wrapped around the inner notes, most commonly containing £20 notes), which are organised into a number of larger bundles, each secured with an elastic band. To analyse these banknotes, the analyst might select a certain number of

Figure 5.2: Photograph taken by MSA Ltd. of a sample of banknotes seized by a law enforcement agency

wraps from the top, middle and bottom of each of the larger bundles (dependent on the total number of banknotes), and analyse these wraps in order, in their entirety. Where samples are analysed in full, they are analysed in order, one bundle at a time. Samples from general circulation are usually analysed in full and in order. As such, the sequential way in which the banknotes are analysed using the mass spectrometer corresponds to the position of the banknotes within the bundle in which they were found. This sequential nature of the data is important when autocorrelation between adjacent banknotes is considered in Section 5.6.2.

Samples of banknotes, both from law enforcement agencies and from general circulation, are taken to the laboratory in sealed and tamper-evident plastic bags. Before opening a sample of banknotes, the forensic analyst takes a number of swabs from the surface being used to lay the banknotes on. Further swabs are taken from the outside of the packaging, to decide if the banknotes could have been contaminated on seizure, the inside of the packaging, and from any extras associated with the samples (e.g. the banknotes could be inside a rucksack). These swabs are tested for traces of drugs in the same way as the banknotes. The measurements obtained from these swabs could be used to determine whether another proposition should be considered in place of H_B, e.g. if the swab taken from the outside of the packaging (which is put in place by law enforcement agencies, on seizure) was contaminated with a large amount of drug, then consideration should be given to the proposition that the banknotes had been contaminated through being placed in a contaminated bag. A photograph of a sample of banknotes, after removal from the tamper-evident bag but before analysis, is shown in figure 5.2. It is clear from the photograph that the sample is arranged in bundles, each secured with a bank strip.

After taking swabs, one or more standards are injected into the mass spectrometer. This is to check that the machine is functioning normally. Standards consist of a liquid solution containing a pre-specified amount of each of the drugs being measured. The banknotes are then analysed, one by one. One half of the banknote, usually the Queen's head end, is held inside the thermal desorption unit for around one second (Dixon et al. (2006); Sleeman et al. (2000)) to give any substances on the banknote the opportunity to evaporate. The forensic analyst waits until the ion count for a banknote

has reduced to below the level of noise before inserting the next banknote into the thermal desorption unit. Any drug ions remaining in the mass spectrometer from the previous banknote could affect the reading on the banknote being currently analysed. This carry-over could lead to autocorrelation between measurements on adjacent banknotes. So, even samples which might not have been analysed in order might still be considered as autocorrelated data.

5.3 The propositions and associated selection of the samples and exhibits for the training data sets

The likelihood ratio is calculated by taking the ratio of the probability density functions of the data under the two propositions: that the banknotes are associated with a person who is involved with criminal activity involving cocaine (H_C), and that the banknotes are associated with a person who is not involved with criminal activity involving cocaine (H_B). Training data are required for H_C and for H_B. These training data can be used with the methods in Chapters 3 and 4 to derive probability density function estimates for the questioned sample, conditional on H_C and on H_B, and hence to make a comparison of the relative values of these density estimates using the likelihood ratio. These sets of training data are denoted C (for H_C) and B (for H_B). Data consisting of samples of banknotes collected from crime scenes and of samples of banknotes taken from general circulation were introduced in Section 5.2.1. Samples collected from crime scenes are those brought in for analysis by law enforcement agencies. Samples from general circulation were collected to compare to those samples collected from crime scenes. For the work in this thesis, a subset of these samples is taken to form the data sets B and C. In this section, the reasoning behind the formation of the two data sets B and C is explained with relation to the two propositions H_B and H_C. It is known that there is some overlap in the quantities of contamination found on those samples in set B and those samples in set C. The problems arising from this overlap will be discussed in Section 5.3.3.

5.3.1 Banknotes that are associated with a person who is involved with drug crime relating to cocaine (data set C)

Samples of banknotes which were suspected of having been involved in crime were received from law enforcement agencies. Each sample is known as an exhibit; a group of exhibits from the same criminal case will be known as a case. The decision of how to divide a case into multiple exhibits is made by the law enforcement agency. As an example, this decision might be made on the basis of the locations in which the samples were found. Different exhibits are brought in for analysis in different tamper-evident bags and they are analysed separately. Exhibits are analysed in a series of multiple runs, where each run is a twenty minute period of analysis on the mass spectrometer. For each exhibit, an expert opinion is formed on whether the exhibit is unusually contaminated or not. This expert opinion may be used as evidence in the criminal case. Feedback is often received from law

enforcement agencies, detailing the outcome of cases (e.g. the result of a trial) in which the banknotes that have been tested have been involved.

In order to construct a database of banknotes which are known to have been associated with a person who has been involved in drug crime involving cocaine, the feedback from the law enforcement agencies was analysed to ascertain which of the exhibits of banknotes were associated with a criminal case in which a suspect was convicted of or pled guilty to a drug crime involving cocaine. These exhibits were used as the training data set C. Set C therefore contains exhibits that were either part of a case which went to trial, and in which a suspect was convicted of a cocaine-related crime, or were part of a case in which a suspect pled guilty to a cocaine-related crime. 29 cases containing at least one exhibit with greater than 20 banknotes were found to fit this description. The 29 cases consist of between one and six exhibits, and there are a total of 70 exhibits which are known to have been associated with a person who was involved in drug crime relating to cocaine. A summary of the exhibits included in the training data set C is given in Appendix D. For future reference, any sample of banknotes used in the analyses discussed that is said to be associated with a person who is involved with drug crime relating to cocaine will be known as an exhibit.

Using the definition of an exhibit given above, the phrase 'the banknotes are associated with a person who is involved with crime involving cocaine' in the context of proposition H_C can be taken to mean that the banknotes have been seized by law enforcement agencies as evidence in a criminal case against a group of one or more people, and that at least one of these people is guilty (in the eyes of the law) of a crime involving cocaine, as this is true for all exhibits in set C. Similarly, a more precise definition of proposition H_B is that the banknotes have been seized by law enforcement agencies as evidence in a criminal case against a group of one or more people, and that none of these people is guilty (in the eyes of the law) of a crime involving cocaine. As discussed in the following section, this latter definition requires an assumption that banknotes from general circulation are representative of banknotes associated with a group of people that would all be found not guilty of a crime involving cocaine. Using these definitions, propositions H_C and H_B are mutually exclusive. However, there is one important limitation. Difficulties arise when there are multiple suspects involved in a criminal case, and banknotes have been seized as evidence against the entire group. A likelihood ratio which is greater than one in this situation only provides support for the proposition that at least one of the suspects from whom the banknotes have been seized is involved with a crime involving cocaine, not that the entire group is involved with a crime involving cocaine. Further limitations relating to proposition H_B are discussed in the following section.

There are two difficulties with the definition of training set C, as described above. The first is that the evidence provided from the forensic analysis and the resulting expert opinion may have influenced the outcome of any trial that has taken place. In fact, just twelve of the 70 crime exhibits in set C were declared as contaminated by the experts, suggesting that this is not too much of an issue. However, this leads to another problem, in that just because the criminal case resulted in a conviction (or a guilty plea), this does not necessarily mean that the banknotes in the individual

exhibits associated with the case were contaminated with cocaine (in fact, 58 of the 70 were not declared by experts as contaminated!) because the exhibits themselves might not have been involved in any cocaine-related activity. These difficulties were further exacerbated by the fact that whilst the result of a case might be known to be a conviction, it is not possible to map this result onto individual exhibits of banknotes. For these reasons, a large number of the exhibits in set C contain banknotes which have contamination which is consistent with banknotes from general circulation. Therefore, it is expected that many exhibits in set C, when treated as the questioned sample, will have likelihood ratios of less than one (and so provide support for H_B). The effect that these difficulties have on the evaluation of the likelihood ratio are discussed further in Section 5.3.3.

There are several possibilities for the improvement of the training data set associated with H_C. In order to resolve the circular problem that the result of a court case might be influenced by the opinion of the expert, a training data set could be used which consists only of exhibits in cases where the suspect has pled guilty. However, this will not completely resolve the problem because suspects might be encouraged to plead guilty to reduce the length of their sentence if evidence against them is overwhelming. Taking this course of action would also reduce the size of the training data set available. Alternatively, surveys in drug clinics could be done. People could be asked, anonymously, if they had used cocaine. If so, their banknotes could be tested and added to the training data set. There are several problems associated with this approach. Firstly, the banknotes obtained would all be from drug users, but not necessarily from drug dealers. Banknotes obtained from drug dealers might have different patterns of contamination to those obtained from drug users, so this approach would not be useful for obtaining evidence relating to drug crimes other than those associated with drug use. Secondly, individuals are unlikely to have (or want to have tested) large quantities of banknotes in their possession when in a drug clinic. All of the exhibits included in the data set C have more than twenty banknotes, and many contain over 100 banknotes. Samples of this size are not likely to be obtained using surveys in drug clinics.

Another approach is to consider exhibits of banknotes that have been involved with crime (rather than exhibits that are associated with a person who is involved with crime). This approach would require an associated change to proposition H_C, but might resolve the problem that many exhibits have contamination consistent with general circulation. To collect data for this new proposition, simulated drug transactions could be carried out in a laboratory using samples of banknotes from general circulation. A problem with this approach is that it is difficult to design an experiment which would accurately reflect a real drug transaction. In reality, there are a large number of ways in which banknotes could be contaminated in the course of drug crime (via use of banknotes directly in drug use such as snorting, via contamination from the hands of people that have been handling drugs, secondary contamination from packaging or tables on which drugs have been stored, etc.). It would therefore be time consuming and expensive to obtain a large database using this approach.

Previous work establishing statistical models for drug traces on banknotes (Jourdan et al. (2013); Besson (2004); Dixon et al. (2006)) have used banknotes seized by law enforcement agencies as set

C, rather than exhibits with the definition seen here. As discussed in Section 2.2.1, there are two problems with using banknotes seized by law enforcement agencies as training data. The first is that there would not necessarily be any association of the banknotes in the training set with a person who is associated with crime, meaning that proposition H_C, used in this thesis (that the banknotes are associated with a person who is involved with crime involving cocaine), could not be used. A proposition stating that the banknotes were seized by law enforcement agencies would have to be used instead. This would not be useful to a judge or jury as it does not help answer any relevant questions about the questioned sample **z**. The second problem is that any seized sample **z** brought in for testing would, by definition, already be known to belong to the set of data associated with proposition H_C. Any model which could discriminate perfectly between a set of banknotes seized by law enforcement agencies and a set of banknotes from general circulation would always assign a seized sample to the former set. The definition of C used here is that of banknotes associated with a case in which the suspect was convicted of a cocaine-related crime, so this problem does not occur.

5.3.2 Banknotes from general circulation (data set B)

Samples of banknotes from general circulation were obtained from a variety of locations around the UK. Information on the currency, the type of place from which the sample was obtained and the location from which the sample was obtained was available for most of the samples. For future reference, any set of banknotes used in the analyses discussed that is associated with general circulation (also known as 'background') will be known as a sample. Samples used to form set B were those that were known to have consisted of UK currency (all exhibits used were also UK currency). No Northern Irish issue banknotes were sampled, so all samples were of English or Scottish currency. In total, 193 samples of banknotes from general circulation were used in the set B. The tables 5.1 and 5.2 show the breakdown of the location and type of place from which the 193 samples of banknotes were obtained. The number of banknotes in each of the samples included in the training data set B is given in Appendix D.

Proposition H_B is that the questioned banknotes are associated with a person who is not involved with criminal activity relating to cocaine. Therefore, to use banknotes from general circulation as the data set B, an assumption has been made that banknotes from general circulation are representative of banknotes that are associated with people who are not involved with criminal activity involving cocaine. This seems like a reasonable assumption to make; in general banknotes associated with people who are not involved with criminal activity involving cocaine are those from general circulation.

As shown in Table 5.1, a large number of general circulation samples were taken from the Avon and Somerset area, highlighting that the samples making up the set B form a convenience sample. Ideally a random stratified sample would be taken, which would reflect the proportions of banknotes in different regions and from different types of location. Tests were carried out in Ebejer, Lloyd et al. (2007) which did not find evidence that the region that banknotes in general circulation are from

has an effect on the quantities of drug found. However, for the calculation of likelihood ratios for questioned samples in a particular case it may be necessary to tailor the general circulation database for use in that case to the region in which the crime occurred, or to refine the analysis with the use of regional factors (though these are not issues that are discussed further here).

Table 5.2 shows that the majority of samples of banknotes were taken from banks. This, again, could be a problem that requires further investigation if the defendant maintained that the banknotes had been acquired in some other way, for example from the sale of a large item such as a used car (which would result in a large transfer of banknotes). A statement from the defendant relating to the acquisition of the banknotes would necessitate a change to proposition H_B. In the case of the sale of a used car, the new proposition could be that the banknotes are associated with a person who has obtained them from the sale of a used car for cash, and who is not associated with drug crime involving cocaine. A database would then need to be set up, consisting of samples of banknotes obtained from the sale of used cars, where the person from whom the banknotes were obtained was not involved with drug crime relating to cocaine. This new database would be used to obtain probability density function estimates for the questioned sample, conditional on the new proposition. Banknotes obtained in such a way may have different contamination patterns to those taken from a bank because banknotes obtained from a bank are more likely to be a mix of banknotes taken from lots of different locations, with lots of previous owners. Banknotes obtained from the sale of a used car could have been contaminated by the previous owner of the banknotes. This change in proposition could therefore have a large effect on the likelihood ratio. Careful consideration of the database B is needed, to match it to the proposition being considered. If the proposition H_B implied that the banknotes had come from a location other than a bank, then the database used here would not be appropriate, due to the large number of banknotes in the database that were obtained from banks.

5.3.3 Problems with the evaluation of the likelihood ratio when training data set C contains a subset which is indistinguishable from training data set B

As discussed in Section 5.3.1, some of the exhibits in set C (banknotes associated with a person who is involved with crime involving cocaine) have contamination consistent with samples from set B (banknotes from general circulation), so that they are indistinguishable from samples in set B. The phrase 'contamination consistent with' is used to mean that some of the exhibits in set C have cocaine measurements which are generated by the same probability density function (with parameters generated by the same between sample distribution) as the cocaine measurements of samples in set B. Set C can therefore be partioned into two subsets: those exhibits that have contamination consistent with samples in set B (C', say), and those that do not have contamination which is consistent with samples in set B (C'', say). It is not known which of these two subsets each of the exhibits in set C belongs to. It is expected that the quantities of cocaine on exhibits in set C'' will be higher than the quantities of cocaine on exhibits in set C'. A description of these two subsets is given by:

Police force area	Number of samples	Police force area	Number of samples
Avon and Somerset	78	Dorset	2
Met	18	Cleveland	2
West Yorkshire	11	Thames Valley	2
Lancashire	9	Humberside	2
Hampshire	6	Strathclyde	2
Northumbria	6	Unknown	2
Scotland (Other/unknown)	6	Berkshire	1
Gloucestershire	4	Northamptonshire	1
Gwent	4	Essex	1
North Yorkshire	4	South Yorkshire	1
Devon and Cornwall	4	West Mercia	1
South Wales	4	Dumfries and Galloway	1
West Midlands	4	Wiltshire	1
Nottinghamshire	3	Aberdeenshire	1
Kent	3	Flintshire	1
Leicestershire	3	Lincolnshire	1
Hertfordshire	2		
Merseyside	2	*Total*	193

Table 5.1: Number of general circulation banknote samples in different police force areas.

Type of location	Number of samples
Bank	157
Newsagent	16
Shopping Centre	4
Bureau de Change	3
Sample provided by police	2
Cash point	2
Post office	1
Sale of car	1
Casino	1
Pub	1
Unknown	5
Total	193

Table 5.2: Number of general circulation banknote samples in each type of location

C' The exhibit has contamination which is consistent with that typically detected on samples from general circulation. This could be because the exhibit was involved in a crime involving cocaine but was not contaminated in the course of it (perhaps no drug was present at the exchange of money), or because the exhibit was not involved in a crime involving cocaine. In this latter scenario, the exhibit may have been obtained innocently, or it may have been involved in some other crime, not involving cocaine.

C'' The exhibit has contamination which is not consistent with that typically detected on samples from general circulation. This could be because the exhibit was contaminated through its use in illegal drug-related activity involving cocaine or in the course of other, legal, drug-related activity involving cocaine. This activity could have been carried out by someone other than the person who was eventually convicted.

These descriptions refer to the banknotes within the exhibits in set C. All of these exhibits were associated with a person who was involved with drug-related crime involving cocaine, whether in set C' or C''. Support for proposition H_C implies support for the proposition that the seized banknotes are associated with a person who is involved with drug-related crime involving cocaine. Support for this proposition does not imply support for the proposition that the seized banknotes themselves were involved with drug-related crime involving cocaine; this is shown by the descriptions of the subsets C' and C'' where in many of the scenarios described, the exhibit was not involved with drug-related crime involving cocaine.

The exhibits in set C' are indistinguishable from samples in set B by definition: exhibits in set C' are defined as being those exhibits with contamination which is consistent with that typically detected on samples from general circulation. The purpose of this section is to describe how methods for the evaluation of the likelihood ratio should be assessed when there is a subset of one training data set which is indistinguishable from the other training data set in this way. The model described here is an 'ideal' model, which discriminates perfectly between samples in B and exhibits in C''. It assumes that a known proportion p of the exhibits in C are in C'. The aim is to evaluate the likelihood ratio for a questioned sample, $\mathbf{z}$, where the likelihood ratio is associated with the two propositions H_C and H_B, as defined previously.

Assume that the contamination of a sample of banknotes in set B can be modelled using the within sample distribution associated with the density function $f(\cdot \mid \theta^B)$, and that the contamination of an exhibit of banknotes in set C can be modelled using the same function, but with a different parameter, given by $f(\cdot \mid \theta^C)$. The parameters θ^B and θ^C vary between samples and exhibits, and so a between sample distribution is used to model this variation. Assume that the between sample density function of θ^B is given by the function $g(\theta^B)$. The between sample distribution of θ^C depends on the variation in the parameters governing the contamination of the two sets, C' and C'', so is more complicated. It is known that exhibits in set C' have contamination consistent with exhibits in set B, so the between sample density function for exhibits in set C' must also be given by the function $g(\cdot)$.

Exhibits in set C'' do not have contamination consistent with samples in set B and exhibits in set C', so the between sample density function for exhibits in set C'' is different. Assume that the between sample density function for exhibits in set C'' is given by $h(\cdot)$. Combining these two functions, the between sample density function for the parameter θ_C is therefore given by the mixture distribution with density function

$$pg(\theta^C) + (1-p)h(\theta^C).$$

The likelihood ratio for a set of questioned banknotes $\mathbf{z}$, analogous to that in (2.3), is therefore given by

$$LR = \frac{\int f(\mathbf{z} \mid \theta^C)(pg(\theta^C) + (1-p)h(\theta^C))\mathrm{d}\theta^C}{\int f(\mathbf{z} \mid \theta^B)g(\theta^B)\mathrm{d}\theta^B}.$$

Splitting the numerator of LR into two integrals, this simplifies [1] to

$$\begin{aligned} LR &= \frac{p\int f(\mathbf{z} \mid \theta^C)g(\theta^C)\mathrm{d}\theta^C + (1-p)\int f(\mathbf{z} \mid \theta^C)h(\theta^C)\mathrm{d}\theta^C)}{\int f(\mathbf{z} \mid \theta^B)g(\theta^B)\mathrm{d}\theta^B} \\ &= p + (1-p)\frac{\int f(\mathbf{z} \mid \theta^C)h(\theta^C)\mathrm{d}\theta^C}{\int f(\mathbf{z} \mid \theta^B)g(\theta^B)\mathrm{d}\theta^B}. \end{aligned} \tag{5.1}$$

Denote by LR' the likelihood ratio given by

$$LR' = \frac{\int f(\mathbf{z} \mid \theta^C)h(\theta^C)\mathrm{d}\theta^C}{\int f(\mathbf{z} \mid \theta^B)g(\theta^B)\mathrm{d}\theta^B}$$

which is seen on the right hand side of (5.1). The likelihood ratio LR' is the likelihood ratio given by $f(\mathbf{z} \mid H_{C''})/f(\mathbf{z} \mid H_B)$, where proposition $H_{C''}$ is the proposition that the banknotes do not have contamination consistent with general circulation and are associated with a person who is involved with crime involving cocaine. An ideal model, which can discriminate perfectly between samples in B and exhibits in C'', would give all samples in B a value $LR' < 1$ and all exhibits in C'' a value $LR' > 1$. This ideal model would therefore also give all exhibits in set C' a value $LR' < 1$ because, by definition, exhibits in set C' are consistent with exhibits in set B, and so the parameters governing the probability density function of exhibits in set C' vary as specified by the between sample density function $g(\cdot)$.

If $LR' < 1$ then the overall likelihood ratio in (5.1), LR, lies between p and 1. As such, if the seized sample $\mathbf{z}$ is from the set C', and if LR' is correctly less than one, then LR lies between p and 1, and so is less than one. This means that even though the model is correctly assigning the value LR', it is providing misleading evidence for $\mathbf{z}$ because the overall likelihood ratio LR is less than one, and hence support is given for proposition H_B, even though $\mathbf{z}$ is in the set C (rates of misleading evidence were introduced in Section 1.1). The proportion of exhibits in set C' is p, so a proportion p

[1] Note that the functions $g(\theta^B)$ and $g(\theta^C)$ are the same function, evaluated at different values. This differs from notation elsewhere in this thesis, e.g. in (2.3), where $f(\theta_1)$ and $f(\theta_2)$ in the numerator and denominator of the likelihood ratio are different functions, evaluated at different values.

of the exhibits will, when treated as the questioned sample, have a likelihood ratio which provides misleading evidence in support of H_B, even when the model is able to discriminate perfectly between the sets B and C''. The proportion p is unknown in practice so, as a result, the rate of misleading evidence for crime exhibits should not be used as a measure of the performance of the model. The rate of misleading evidence for general circulation samples may, however, still be used because if the seized sample $\mathbf{z}$ is in B, then a perfectly discriminating model will assign $LR < 1$.

If $LR' > 1$, then the overall likelihood ratio LR is greater than one. As such, if the seized sample $\mathbf{z}$ is from the set C'' and LR' is correctly greater than one, then LR is also greater than one. This means that the likelihood ratio when exhibits in set C'' are treated as the questioned sample, does correctly give support to proposition H_C, that the questioned sample is in set C.

As an aside, also note that (5.1) can be rearranged to give

$$LR' = \frac{1}{1-p}(LR - p). \tag{5.2}$$

If $LR < 1$ then LR' is also less than one. If $LR > 1$ then LR' is also greater than one and will always be greater than or equal to LR (note that, from (5.1), $LR > p$). So, a likelihood ratio that supports proposition H_C, that the banknotes are associated with a person who is involved with crime involving cocaine, over the converse, H_B, will also support the proposition $H_{C''}$, that the banknotes do not have contamination consistent with general circulation and are associated with a person who is involved with crime involving cocaine, over H_B.

5.4 The peak detection algorithm

5.4.1 Introduction

The data obtained in Section 5.2.1 consisted of gas phase ion transition counts per second, measured ten times per second, for a sample of banknotes. There are multiple measurements for each banknote, and information detailing which measurements correspond to which banknotes is not available. In this section, an algorithm developed to convert the raw ion transition counts into a set of 'peak areas' is described. One peak area is calculated per banknote, and a peak area corresponds to the sum of the ion transition counts per second that have been measured for a particular banknote. The peak area of a banknote gives a measure of the quantity of drug on that banknote.

The Mass Spectrometer output is recorded in a so-called wiff file (or for older data, a netcdf file), and takes the form of the gas phase ion transition count per second at each scan number, for either eight or ten ions (see Section 5.2.1 for details). The scan number does not correspond to the banknote, but to the time. As a banknote is passed through the heated plates of the thermal desorption unit, measurements of the ion transition count are carried out for each ion transition in turn. As substances on the banknote are evaporated, the ion transition count eventually reduces to a level around the noise threshold (determined by visual inspection by the analyst), and then the next banknote is

passed through the heated plates. The data for each ion transition therefore resemble a series of peaks, with each peak corresponding to a banknote. Examples of the outputs coming from a crime exhibit (involving cocaine) and a general circulation sample are provided in figures 5.3 and 5.4 respectively. Both of these figures shown ten ion transitions. The only two that can be seen are the two cocaine ion transitions (304 → 182 in red and 304 → 105 in blue). The ion counts of the other eight ion transitions are too small to be seen. The large peak at around the eight minute point in Figure 5.3 shows an example of the ion count not fully reducing to below the noise threshold before the next banknote is inserted into the thermal desorption unit of the mass spectrometer. As described in Section 5.2.1, ion transitions from the banknote inserted at around the eight minute point may have been counted towards the ion transition count of the next banknote. This carry-over could be a source of the autocorrelation between the measurements of adjacent banknotes.

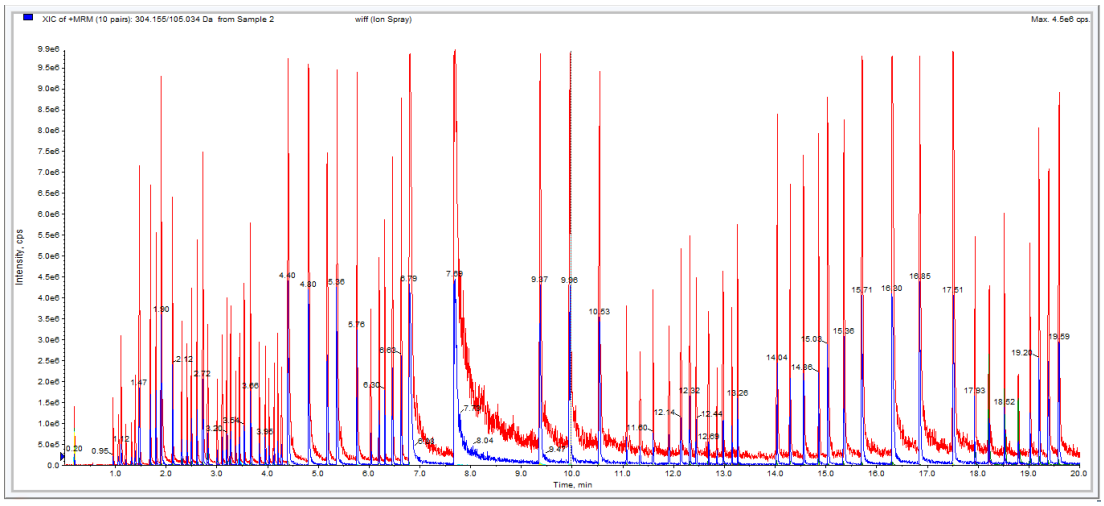

Figure 5.3: Ion transition counts per second for one run from an exhibit in a criminal case. Ion transition 304 → 182 is shown in red and ion transition 305 → 105 is shown in blue.

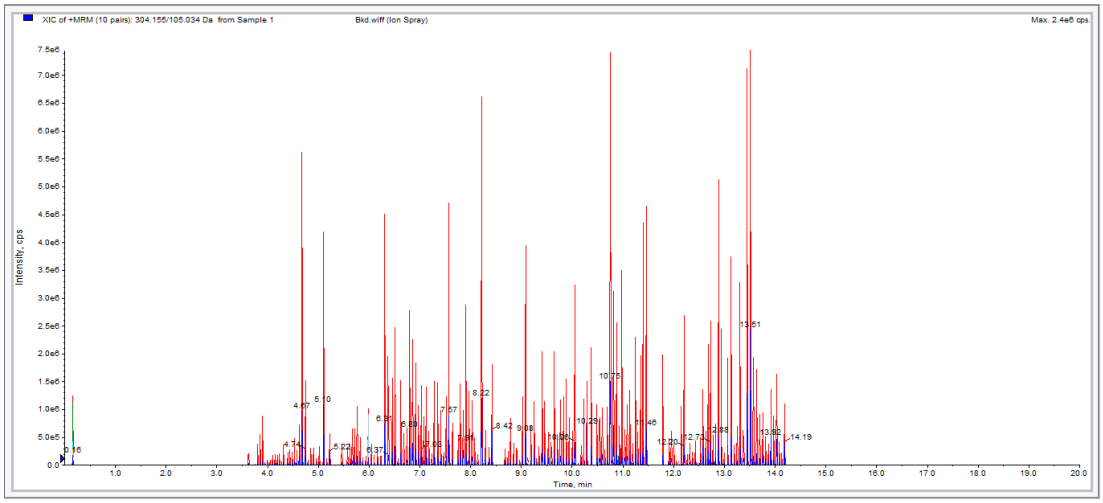

Figure 5.4: Ion transition counts per second for a general circulation sample. Ion transition 304 → 182 is shown in red and ion transition 305 → 105 is shown in blue.

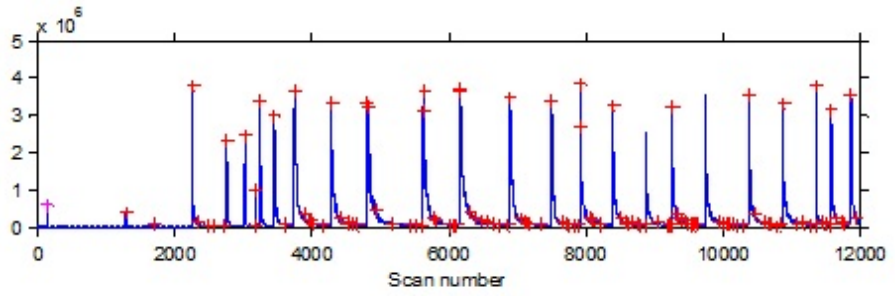

Figure 5.5: Detected peaks in a crime exhibit run, using the algorithm and graphical user interface given in Lloyd (2009)

Raw data recorded in a wiff file were converted to a mat file (Matlab), using code provided by Gavin Lloyd (Lloyd (2009)). Mat files could then be read into R using the R.Matlab package (Bengtsson (2013)). Each wiff file consisted of multiple runs of data, with each run belonging to the same case. A run corresponds to a twenty minute period of analysis on the mass spectrometer. Large cases were spread over several wiff files. Netcdf files consisted of just one run each, and were read directly into R using the ncdf package (Pierce (2011)). The following data were extracted from each file (wiff and netcdf):

- Number of runs.
- Number of different ion transitions for which the data were collected, and mass to charge ratio, *m/z*, of each of the precursor and product ions.
- Name of case and name of each run.
- Intensity (gas phase ion transition count per second) at each scan number for each run and each ion transition.

In order to calculate the peak areas, the scan numbers for the start and end of each peak must be detected. These scan numbers correspond to the time at which the banknote was inserted into the thermal desorption unit, and the time at which it was removed. The aim of the peak detection algorithm given here is to find these scan numbers. This algorithm contains some steps from the algorithm developed first in Dixon et al. (2006) and adapted later in Lloyd (2009), which uses the change in derivative to identify the start and the end of the peak. The algorithm in Lloyd (2009) was developed for samples of banknotes taken from general circulation, which in general have less contamination and hence less noise than those that have been involved in crime. As a result, when this algorithm was applied to crime exhibits, it tended to detect peaks where there were no banknotes. Figure 5.5 gives an illustration of the results obtained when the algorithm given in Lloyd (2009) was used on a run from a crime exhibit. Detected peaks are shown with red crosses. As can be seen, there are many peaks detected which do not correspond to banknotes.

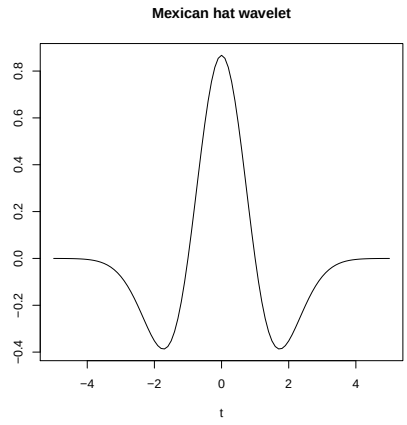

Figure 5.6: The Mexican hat wavelet

5.4.2 The MassSpecWavelet algorithm

The algorithm developed for use in this thesis consists of three parts: the location of peaks, the removal of peaks falsely identified as banknotes, and the determination of the start and end of the peak. The first step uses Du, Kibbe and Lin's MassSpecWavelet algorithm (Du et al. (2006)), which fits 'Mexican Hat' wavelets (figure 5.6) at each scan number for a variety of different scales, using the Continuous Wavelet Transform (CWT). The CWT is given in Du et al. (2006) by:

$$C(a,b) = \int_{\mathbb{R}} s(t)\psi_{a,b}(t)\,dt \tag{5.3}$$

where the wavelet $\psi_{a,b}(t)$ is defined as

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right),\ a \in \mathbb{R}^+ - \{0\},\ b \in \mathbb{R}. \tag{5.4}$$

Here, $s(t)$ represents the signal (which is the ion transition count at scan t), a is the scale of the wavelet and b is the translation. The wavelet $\psi(t)$ is given by the Mexican Hat wavelet, so that

$$\psi(t) = \frac{2}{\sqrt{3}}\pi^{-\frac{1}{4}}\left(1-t^2\right)e^{-\frac{t^2}{2}}.$$

Using the MassSpecWavelet algorithm, the matrix $C = \{C(a,b)\}$ is constructed by calculating $C(a,b)$ in (5.3) at different translations (b) and scales (a). The larger the value of $C(a,b)$ for a particular a and b, the better the wavelet with coefficients a and b matches the signal $s(t)$. Assuming that the peaks in the signal $s(t)$ are symmetric and each has one maximum, for a fixed scale a, $C(a,b)$ is locally maximised when b is equal to the scan number at the centre of the peak (the assumed maximum of the peak). This is because the Mexican Hat wavelet $\psi(t)$ has its maximum at $t = 0$. As explained in Du et al. (2006), when b is fixed at the value that gives this local maximum, the value of $C(a,b)$ is maximised when 'the scale best matches the peak width'. This behaviour results in ridges in the three-dimensional plot which has values of a and b forming the x and y co-ordinates, and the value

of $C(a,b)$ giving the height of the plot. The best fitting wavelets can therefore be found by finding the positions of the ridges in the matrix C. The values of b at these ridges will then correspond to the times at which the peaks occurred, and hence the scan numbers of the peak maxima. The algorithm for finding the positions of the ridges is given in Du et al. (2006).

As discussed in Lloyd (2009), the majority of banknotes, even those from general circulation, have traces of cocaine present on their surface. Therefore, as also done in Lloyd (2009), the positions of peaks for the two cocaine ion transitions were used to determine the positions of the peaks for the remaining ion transitions (where there may be no peak). The ion counts for the two cocaine ion transitions were summed together and used as the signal $s(t)$ in the algorithm, to increase the size of the peaks. To reduce the noise, the signal was smoothed by replacing the value of each data point by the average of it and the values of the two adjacent data points. The MassSpecWavelet algorithm, from the MassSpecWavelet package in R (Du et al. (2006)) was then run on these data as the first step of the algorithm. The MassSpecWavelet algorithm returns a vector of peak indices, corresponding to the estimated scan number of the maximum of each of the peaks.

The MassSpecWavelet algorithm is designed for Mass Spectrometry data relating to the mass spectrum, e.g. data which take the form of an ion count for different mass to charge, m/z, ratios. The data here are not of this form, as they measure the ion count over time (at different scan numbers). The assumptions of a symmetric and unimodal peak shape do not always hold for this type of data (see Dixon *et. al* (2006) also). As such, whilst the MassSpecWavelet algorithm was very good at picking out all of the peaks, there were often a large number of peaks detected where there were no banknotes, due to the multimodal and asymmetrical nature of many of the peaks. Further steps were developed to remove those peaks which were falsely identified as banknotes; this is the next stage in the algorithm.

5.4.3 The detection and removal of peaks falsely identified as banknotes

The positions of the peaks were taken to be the peak indices returned by the MassSpecWavelet algorithm. Figure 5.7 gives examples of some peaks which were identified by the MassSpecWavelet algorithm but that did not correspond to banknotes. Peaks which do correspond to banknotes are generally fairly evenly spaced throughout the run, so peaks which are much closer together than other peaks within the run are likely to be falsely identified. The graph on the left hand side of figure 5.7 corresponds to one banknote, but contains two peaks. The centre graph shows increased noise levels following the insertion of a series of contaminated banknotes; three peaks are detected where there are no banknotes. The graph on the right shows a jagged peak with large amounts of noise in the descent. Four peaks were detected for one banknote.

A series of rules were devised to remove or merge peaks, and to locate the standards, injected at the start of runs. These rules were designed to merge peaks based on: the distance between them, the dip between them and the second difference between peak positions. In addition, peaks which were small in comparison to an estimated noise line were removed. In cases where peaks were merged, the

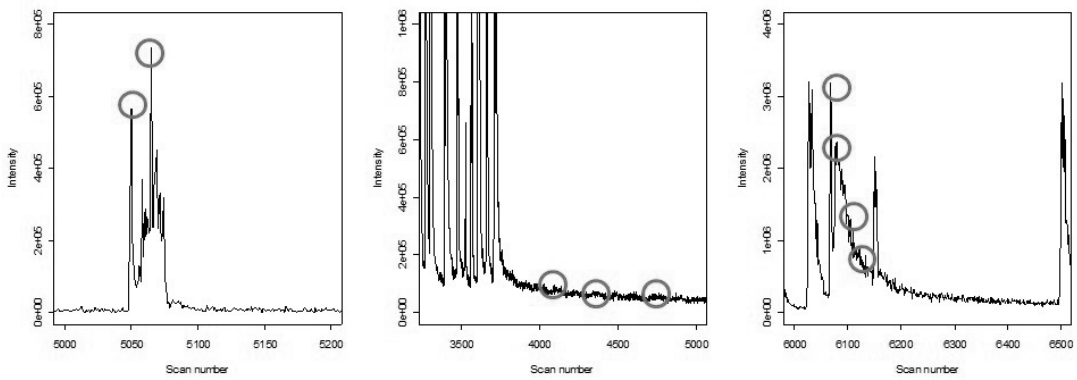

Figure 5.7: Examples of peaks detected by the MassSpecWavelet algorithm that did not correspond to banknotes

peak index chosen to replace the two merged peaks was the peak index of the higher of the two peaks. The following steps were carried out:

1. All peaks within 10 scans of each other were merged.

2. All peaks for which the 25% quantile of the intensities between the two peak positions was bigger than the minimum of the two peaks multiplied by (0.2 + the distance between the peaks/100) were merged. This step merges peaks with an insufficiently large dip between them, with the required dip increasing as the distance between the peaks increases. Falsely identified peaks such as those on the right hand side of figure 5.7 were removed in this way. This step, and the previous step were carried out starting with the pair of eligible peaks with the smallest 25% quantile of intensities between them, with all of the peak indices updated after each merging had occurred.

3. The differences in positions between all pairs of peaks were calculated. Peaks which did not correspond to banknotes were often found to be very close to peaks which correctly identified banknotes (see e.g. the left and right hand diagrams in figure 5.7). The distance between a correctly identified peak and an incorrectly identified peak was small in comparison to the difference between two peaks that correctly identified banknotes. To locate these falsely identified peaks, the differences in peak positions were sorted, and the differences between these differences were calculated. To ensure that fewer than 25% of the peaks were considered outliers, and so that very large peak differences were not considered (caused by breaks in the analysis), the largest of the differences in the first 25% of differences of sorted differences was selected. If this selected second difference was smaller than a threshold (the 75% quantile of the second differences + 2 × the interquartile range of the second differences) then no action was taken. If however, the selected second difference was higher than this threshold, then the smaller of the two differences making up this second difference was selected. Each difference corresponds to a pair of peaks. All pairs of peaks with a difference equal to or smaller than the

selected difference must be unusually close together (since there was a jump in the second difference). Therefore, all of the pairs of peaks with a difference equal to or smaller than the selected distance were merged, as long as this difference was smaller than a threshold of 20. This step was not used for runs with fewer than ten peaks.

For example, let $d_{(1)}, d_{(2)}, \ldots, d_{(n-1)}$ be the $n-1$ sorted differences between n detected peaks. Let $d_{12}, d_{23}, \ldots, d_{n-2,n-1}$ be the second differences between these sorted differences. The largest of the first 25% of these second differences was selected. Assume that d_{23} was the largest such second difference and hence was the selected second difference. If d_{23} was greater than the threshold described above, then the pairs of peaks corresponding to $d_{(1)}$ and $d_{(2)}$ would be merged, as long as each was smaller than 20. The large value of the second difference d_{23} implies that there is a jump between the differences $d_{(2)}$ and $d_{(3)}$, so that differences of size $d_{(2)}$ and smaller are unusually small.

4. An initial noise line was calculated, using a method closely based on that used in Lloyd (2009). Windows were created with 100 scans in each window. If a window had a peak within it or within 20 scans either side of the window, the window was classified as a peak window. Points were placed at the midpoint of each window, with an intensity of the 10% quantile of the intensities within the window for peak windows, and the 90% quantile for non-peak windows. A horizontal line was then fitted through these points, using the least squares method, after discarding those points with intensities lying outside the interquartile range.

5. Pairs of peaks for which the intensities between them did not dip beneath the noise line calculated in step four, and which had an average intensity between peaks which was larger than the minimum peak of the two being considered were merged. This merging was done beginning with the largest scan number. This step was designed to remove falsely identified peaks which corresponded to noise on a correctly identified peak, but where the falsely identified peak and the correctly identified peak were too far apart to be merged using step three, such as those seen on the right hand side of figure 5.7.

6. Peaks which had a peak maximum below twice the noise line were removed (checking for larger peak maxima two scans either side of the currently detected peak position, in case the peak maximum position had not been correctly identified).

7. To locate the standard injections (which vary in number for each run), a method similar to that used in step three, based on second differences, was used. Standards are generally followed by a break in the analysis, indicating that a jump in the difference between peak positions should occur. The first three second differences in the run were calculated. The last such second difference which was above a threshold was found (larger than 200, and larger than the 75% quantile of the second differences excluding the first three, added to 10 x the interquartile range). All peaks occurring before this jump in the second difference were taken to be standards. For

some samples (particularly crime exhibits), the approximate positions of the standard injections were known. For these samples, all detected peaks with a scan number smaller than halfway between the approximate position of the last standard and the first banknote were taken to be standards.

After performing these peak merging and removal steps, and locating the standards, the next stage was to calculate the start and end of each of the detected peaks. This was done using the following steps:

1. The noise line was refitted to the new peak indices, using the method described previously in step 4.
2. The initial value for the start of each peak was taken to be the first point to the left of the detected peak position (usually the maximum) which was both below the noise line and which had another point within 3 scans which was also below the noise line.
3. If this point did not exist, the minimum point between the peak and the nearest peak to the left was used, or the minimum point between the peak and the start of the data.
4. The initial value for the end of the peak was taken to be the first point to the right of the peak position which was both below the noise line and which had another point within 3 scans which was also below the noise line.
5. If this point did not exist, the minimum point between the peak and the nearest peak to the right was used, or the minimum point between the peak and the end of the data.
6. The noise line was refitted using the estimates for the start and end of each of the peaks (instead of the peak positions). Points were placed at the midpoints of each window, with intensity equal to the proportion of the window in a peak multiplied by the 10% quantile of the within peak window intensites, added to the proportion of the window not in a peak, multiplied by the 90% quantile of the window intensities not in peaks. A horizontal line was then fitted through these points, as before (using least squares).
7. Peak starts and ends were then recalculated using the new noise line and the method given in steps 2-5.

The scan numbers of the starts and ends of each peak were taken to be the same across all of the ion transitions measured for the run. Peak areas for each ion transition could then be calculated as the sum of the intensities (ion counts per second) between the start and end of each peak, and the peak maxima could be calculated as the largest intensity between the start and the end of each peak.

Figures 5.8, 5.9 and 5.10 illustrate the results for one run in a particular crime exhibit, after certain stages of the peak detection algorithm have been implemented. The top diagram in figure 5.8 shows the initial peak indices that were detected by the MassSpecWavelet algorithm. As can be seen,

there are many peaks falsely identified as banknotes. The bottom diagram shows the peak indices after carrying out steps one and two of the steps for peak merging and removal. As can be seen, all of the falsely identified peaks that were detected due to multimodality of the peaks have been removed. In the top diagram of figure 5.9, the first peak (a standard) has been identified and removed. In this example, no falsely identified peaks were removed based on the second differences, as there are the same number of falsely identified peaks in the top diagram of figure 5.9 as in the bottom diagram of figure 5.8 (aside from the standard). The peak indices after the removal of peaks using stages 5 and 6, based on the noise line, are shown in the bottom diagram of figure 5.9. These stages remove small falsely identified peaks, caused by noise, with stage 5 designed to remove small falsely identified peaks in the tails of larger peaks. In the bottom diagram of figure 5.9, all of the falsely identified peaks have been removed; the 20 peaks correspond to 20 banknotes. Figure 5.10 shows the final results obtained for two of the ion transitions. The peak maxima detected are marked with a red circle, the start of the peak with a red cross and the end of the peak with a green square.

In figure 5.11, the results for another run in the same crime exhibit are illustrated. In the top diagram, the initial peak indices detected by the MassSpecWavelet algorithm are shown. In the bottom diagram, the peak indices after stages 1-7 of the peak merging and removal steps have been implemented are displayed. The large peak seen just after scan number 2,000 has not been identified as a peak because it corresponds to a swab, not a banknote (for more details on swabs, see Section 5.2.3). The runs shown in figure 5.11 and figure 5.5 are the same. In figure 5.5, the peaks have been detected using the algorithm developed in Lloyd (2009). As can be seen, the peak detection algorithm developed for use in this thesis has correctly identified and removed the peaks that were falsely identified as banknotes in figure 5.5.

The peak detection algorithm described in this section was implemented for all of the runs in all of the samples and exhibits selected for the training data sets B and C. For the work in this thesis, only the peak areas relating to the ion transition $304 \rightarrow 105$ were used. Using the information found in case files, the peak areas for individual runs were collated and formed into exhibits (there are often multiple runs in each exhibit). General circulation samples were usually analysed in one run. In Dixon et al. (2006) and Lloyd (2009), the log-transformed peak areas (to base ten) were assumed to be Normally distributed. As the autoregressive model and the hidden Markov model both assume Normality for the error terms, the peak areas (for both general circulation samples and crime exhibits) are log-transformed (to base ten) here also. This transformation reduces positive skew and improves the fit of the Normal distribution to the data. Therefore, the logarithms of the peak areas calculated using the algorithm described here are the data analysed.

5.5 Data validation and verification

Various error checks were carried out on the data. The total number of banknotes in each general circulation sample, and the total number of banknotes in each crime exhibit run were available.

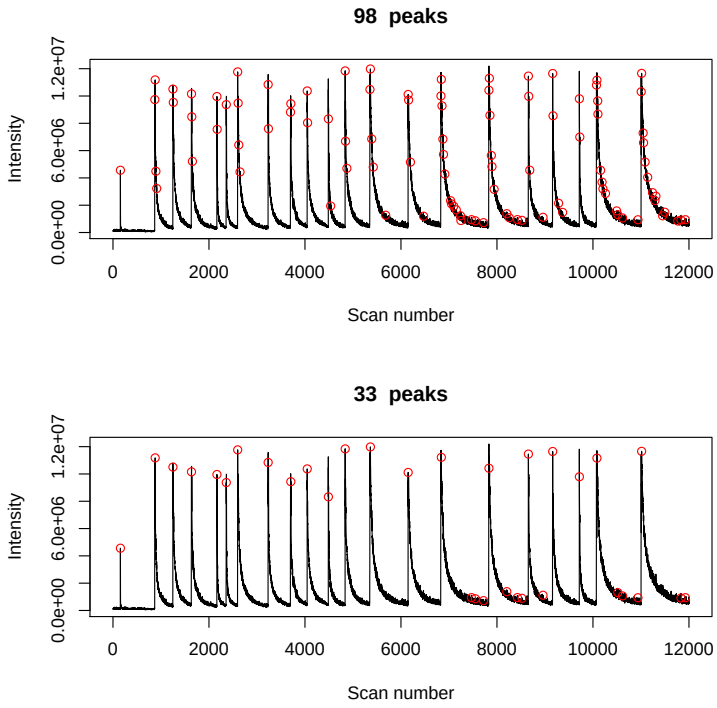

Figure 5.8: Inital peak indices (top) and peak indices after stages 1 and 2 (bottom) of peak merging. A detected peak is indicated with a circle.

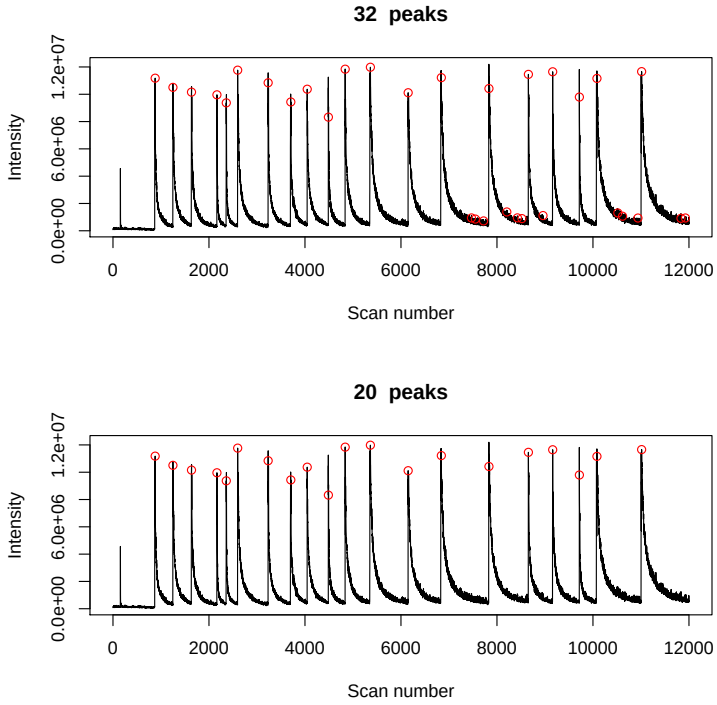

Figure 5.9: Peak indices after stage 3 (top), and after stages 5 and 6 (bottom) of peak merging.

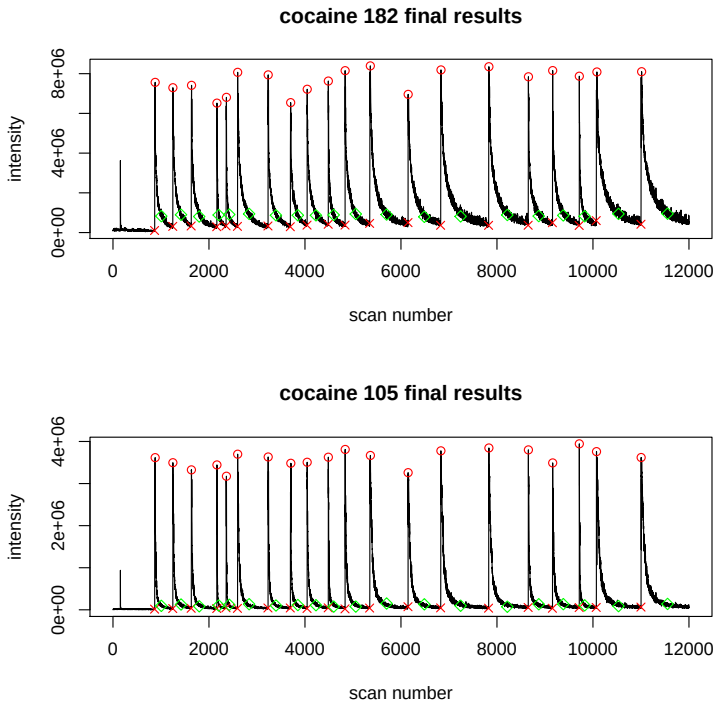

Figure 5.10: Start and end points of peaks, illustrated for the two cocaine ion transitions

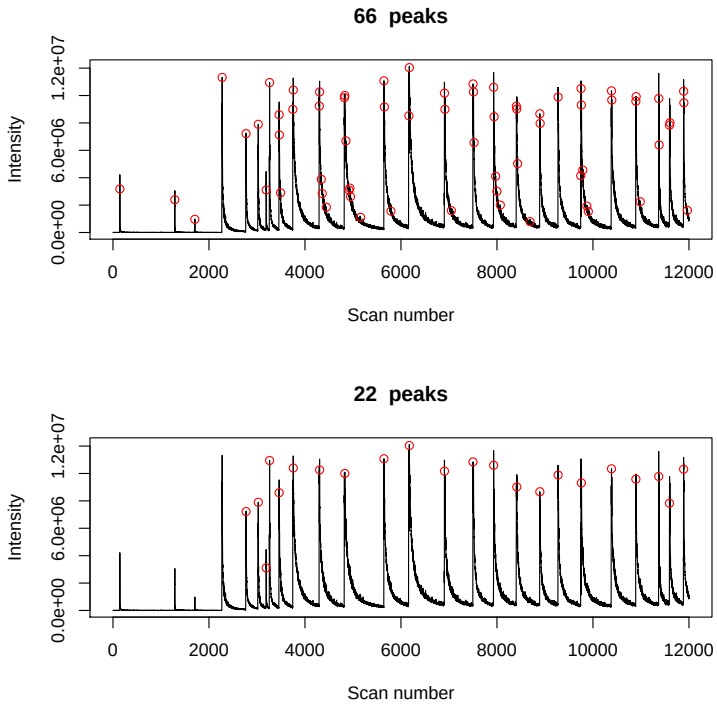

Figure 5.11: Initial peak indices (top) and peak indices after all stages of peak merging have occurred (bottom). This is the same run as seen in figure 5.5

These numbers were checked against the number of peaks detected by the peak detection algorithm (which should correspond to the number of banknotes), and the difference between the number of peaks detected and the true number of banknotes was calculated. Any sample or exhibit for which this difference was greater than 10% (either way) of the correct value was discarded, because a large difference could imply a poor performance from the peak detection algorithm. Any general circulation sample or crime exhibit with fewer than twenty detected peaks in total was not used in the analysis. In general, exhibits with fewer than twenty banknotes are not analysed by forensic experts due to the difficulty of drawing conclusions from such small samples. In addition, particularly for the hidden Markov model, the number of observations in each sample must be sufficient to obtain good parameter estimates.

For some exhibits, individual runs were missing. For these exhibits, the longest continuous section of the exhibit was included, with the rest discarded, e.g. if an exhibit consisted of 4 runs, each with 100 detected peaks, and the third run was missing, then the first and second runs would be combined to form the exhibit. This step was taken to maintain the sequential nature of the data. Trace plots of the data were examined, and any outlying data points were further investigated. As a result of these investigations, five detected peaks were removed from the end of one exhibit (exhibit 39) because they were found to be incorrectly identified peaks, which did not correspond to banknotes.

For general circulation banknotes, any samples for which the information on the currency or the total number of banknotes was not available were removed from the analysis. General circulation samples were often tested twice, once for each end of the banknote. The data used here arose from only one of these tests, so that each sample of banknotes only appears in the training data set once. In each run, a standard was injected at the start of the analysis. For crime exhibits, the approximate scan numbers of the standard injections were available; this was not the case for general circulation samples. For general circulation samples, the scan numbers of any standard injections were estimated, as described in the description of the peak detection algorithm (Section 5.4). Graphs of the detected peaks were inspected and any peaks which had obviously been misidentified as standards (seven in total) were reassigned as peaks.

As discussed in Section 5.2.3, swabs were sometimes taken of packaging associated with the banknotes. These swabs were tested for traces of drugs at the same time as the banknotes, and using the same method and equipment. Most swabs were tested at the start of the run, and were therefore treated as standards by the peak detection algorithm (and subsequently removed from the analysis). For some crime exhibits, these swabs were known to have been tested part way through the run. The approximate scan numbers of these swabs were known. Any peaks identified within five scans of the approximate position of a swab were removed from the analysis as the peak was likely to have corresponded to the testing of a swab rather than a banknote.

As discussed in Section 5.2.1, the laboratory has four different machines on which the banknotes may be analysed. The choice of which machine to use in any given situation was made for operational reasons. To take into account differences in response between machines, the peak area measurements

were standardised. The standard injections for the 193 general circulation samples were analysed, and where this standard could easily be identified (for 162 samples), the peak area was calculated. There were often multiple standard injections and swabs for each run, so it was not always clear which peak corresponded to the standard injection of a known amount of solution. An average standard peak area for each of the four machines was calculated based on these 162 samples, and the ratio of each of the average standard peak areas to the average standard peak area of the first machine (machine A) was calculated. The peak areas for each crime exhibit and general circulation sample were then divided by the appropriate ratio, based on the machine used to do the analysis.

5.6 Data exploration

The data consist of the set B, the logarithms of the peak areas of 193 samples of banknotes from general circulation ($\mathbf{x}$) and the set C, the logarithms of the peak areas of banknotes from 70 crime exhibits ($\mathbf{y}$). The data relate to the cocaine product ion, m/z 105. Three aspects of the data are considered. The first is the extent to which banknotes are contaminated with cocaine. The second is autocorrelation. The final consideration is whether different bundles within a sample or exhibit of banknotes, with each bundle potentially coming from a different location, could have different levels of contamination. These properties were the motivation behind the development of the models in Chapter 3.

5.6.1 Cocaine contamination on banknotes

Summary

Figures 5.12, 5.13 and 5.14 show plots of the means and ranges of all 70 crime exhibits and 193 general circulation samples. One general circulation sample, the 158th, had one banknote with a log peak area of 2.46. This outlier was not included in the plot, so that the means and variances of the remaining samples could be seen more clearly. The exhibits in figure 5.12 are arranged roughly by case, meaning that exhibits from the same case are generally grouped together [2]. Exhibits from the same case are more likely to have similar quantities of contamination, which is why samples that are close together in the plot tend to have more similar means and ranges than those further apart. As can be seen in figure 5.12, most banknotes in set C (associated with a person who is involved with crime relating to cocaine) have log peak areas ranging from around 5 to around 8.5. Most banknotes in set B (general circulation), have log peak areas ranging from around 4.5 to around 8. There is a substantial overlap between the ranges occupied by banknotes from the two sets. In addition, the means and the ranges vary a lot within the sets B and C. This is particularly true for crime exhibits; there are exhibits in set C with ranges of contamination that do not overlap with the ranges of contamination of other exhibits in set C.

[2] A table indicating which exhibits are in the same case is given in Appendix D

Cocaine contamination on general circulation banknotes

As discussed in Section 2.2.2, using the percentage of contaminated banknotes within a sample as a statistic to distinguish between general circulation samples and crime exhibits is not possible due to the high frequency of contamination within samples from general circulation. Experiments (Carter et al. (2003)) have shown that 60% of samples of uncontaminated paper swabs which were sent to banks were found to be contaminated with cocaine to some degree after having been counted by the banks (mainly through counting machines). This could go some way to explaining the contamination found on general circulation banknotes. As a result it is not sufficient to focus on the proportion of contaminated banknotes: the quantity of contamination needs to be taken into account to differentiate between crime exhibits and general circulation samples. The peak area is a measure of this quantity of contamination. Figure 5.15 shows the density plot of the mean of the log peak areas of the 193 general circulation samples (dashed line) against the mean of the log peak areas of the 70 crime exhibits (solid line), where the means are determined from the individual log peak areas of the banknotes in the separate samples and exhibits. As can be seen, general circulation samples have mean log peak areas which, although generally lower than those of crime exhibits, are still high, and there is a substantial overlap between the two density plots. This overlap in log peak areas can also be seen in the plots in figures 5.12, 5.13 and 5.14. In addition, in figure 5.15 it can be seen that the density plot of the means of the crime exhibits has a mode at around 6.5, in a similar position to the main mode of the density plot of the means of the general circulation samples, as well as a mode at about 7.2. This is further evidence to suggest that the set C (associated with a person who is associated with cocaine related crime) contains a large number of exhibits which have contamination consistent with samples from general circulation (and so are in set C'), as discussed in Section 5.3.3.

5.6.2 Autocorrelation

Experiments carried out in Ebejer, Winn et al. (2007) indicated that it was possible for drug traces to pass from one contaminated banknote to an adjacent one (although these experiments were for heroin rather than cocaine). As described in Section 5.2.3, general circulation samples and smaller crime exhibits are usually analysed in their entirety and in the order in which the notes occur in the sample or exhibit. For larger exhibits, a proportion of the banknotes are analysed, but they are usually analysed in pre-selected groups of banknotes, where the banknotes in each group were adjacent in the exhibit. As a result, any transfer of cocaine that had occurred between adjacent banknotes in the sample or exhibit, as discussed in Ebejer, Winn et al. (2007), would result in autocorrelation being present within the analysed sample or exhibit. In addition, as also discussed in Section 5.2.3, when banknotes are analysed, there is often no definitive end to the mass spectrometry peak, as shown in figure 5.3. Some of the ion transition counts arising from cocaine on a banknote may be included in the peak area associated with the next banknote analysed. This effect could also result in autocorrelated data.

To check for autocorrelation, the sample autocorrelation coefficient was calculated for each sample and exhibit. It was found that 89% of the 193 samples from general circulation and 77% of the 70 crime exhibits had significant autocorrelation at lag one. A sample or exhibit consisting of n banknotes is said to have significant autocorrelation at lag one if the observed autocorrelation coefficient at lag one is outside the approximate 95% probability interval given by $(-2/\sqrt{n}, 2/\sqrt{n})$, as given on p. 56 of Chatfield (2004). This probability interval is based on a null hypothesis that the autocorrelation coefficients at all lags are zero. The proportions of samples and exhibits with significant autocorrelation at higher orders dropped to 62% and 56% for samples and exhibits, respectively, at lag two, and 35% and 39% by lag five. Any model used for these data should therefore take autocorrelation into account.

The sample partial autocorrelation function was also calculated for each of the samples and exhibits. The partial autocorrelation function at lag p determines the correlation that was not accounted for by an autoregressive model of lag $p-1$. It was found that 89% of general circulation samples and 77% of crime exhibits had a significant sample partial autocorrelation coefficient at lag one at the 5% significance level (where the definition of significant is the same as for the autocorrelation coefficient); these numbers are the same as those for the sample autocorrelation coefficient at lag one, which is expected because at lag one the two coefficients are equal. The number of samples and exhibits with significant partial autocorrelation coefficients dropped to 24% of general circulation samples and 14% of crime exhibits at lag two and 6% of samples and 7% of exhibits at lag five. The large number of samples and exhibits with significant partial autocorrelation coefficients at lag one and the much smaller number with significant partial autocorrelation coefficients at larger lags implies that an autoregressive model of order one would be appropriate for these data (see p62 of Chatfield (2004)). As such, in the following chapter, models taking lag one autocorrelation into account will be fitted to these data. Any autocorrelation at higher lags is not modelled.

5.6.3 Differences in contamination on different bundles of banknotes within the same exhibit or sample

Figure 5.16 shows an estimate of the conditional density function (computed using the np package in R (Hayfield and Racine (2008))) of the log peak area of one banknote conditional on the log peak area of the previous banknote, for one of the general circulation samples. The conditional density function looks like a typical autoregressive process with positive autocorrelation; the cross-section at each value of the previous banknote is unimodal and the higher the log peak area of the previous banknote, the higher the expected log peak area of the following banknote. Similar plots of other samples and exhibits did not all have this unimodal cross-section. Whilst some did resemble an autoregressive process, others had many ridges and had some multimodal cross-sections. This was particularly true for crime exhibits, which could reflect that a sample of banknotes taken from someone involved in drug crime might be a mixture of banknotes which are highly contaminated, and of banknotes

which have come from general circulation. For example, a drug dealer might have acquired bundles of banknotes from many different customers. Some of these bundles might have been contaminated from the customer's drug use, and some of these bundles might be from general circulation. An exhibit seized from this drug dealer would therefore consist of many bundles of banknotes, some of which would have high levels of contamination, and some of which would have low levels of contamination. On the other hand, if a drug dealer were to contaminate an exhibit evenly himself (say, by counting out banknotes onto a contaminated surface), or if he had obtained all of the bundles from the same customer, then the analysed exhibit might consist of similarly contaminated banknotes throughout. Some general circulation samples also exhibited multimodality in the cross-sections of the conditional density functions, although these were fewer in number than for the crime exhibits. This could be because a general circulation sample might still contain bundles of banknotes which have come from different locations and hence have different levels of contamination.

Figure 5.17, shows estimates of the conditional density functions of the log peak area of a banknote given the log peak area of the previous banknote, for three of the crime exhibits. The top plot, exhibit 8, is the conditional density function of an exhibit with 71 detected banknotes. The plot has two ridges, one large and one small. This indicates that a banknote could be followed either by a banknote with contamination comensurate with the larger ridge, or by a banknote with contamination comensurate with the smaller ridge. An autoregressive process with two mean levels, with the mean determined by a latent state can be used to model this behaviour.

The central plot (exhibit 4) shows the conditional density plot of another exhibit, this time with 132 banknotes. The relative weights of the ridges in this plot change as the log peak area of the previous banknote changes. This plot has multimodal cross-sections, as with exhibit 8, but the cross-sections look very different when the previous banknote has a low log peak area, compared to when the previous banknote has a high log peak area. When the previous banknote has a low log peak area, the current banknote is most likely to be in the lower ridge, whereas when the previous banknote has a high log peak area, the current banknote is most likely to be in the highest ridge. This indicates that banknotes with a low log peak area are likely to be followed by another banknote with a low log peak area, and banknotes with a high log peak area are likely to be followed by another banknote with a high log peak area. This dependence, of whether the banknote is 'low' or 'high' on the log peak area of the previous banknote, can be modelled using an autoregressive process with mean determined by a latent state, where the latent states form a Markov chain (i.e. a hidden Markov model), so that high log peak areas are more likely to follow other high log peak areas and low log peak areas are more likely to follow other low log peak areas. Contamination of this pattern would occur if the exhibit contained multiple bundles of banknotes from different origins, with banknotes within the same bundle remaining together within the exhibit. So, rather than banknotes that are contaminated at a high level being randomly placed throughout the exhibit, they are instead grouped together in bundles. The probability of a banknote being contaminated at a high level depends on the level of contamination of the previous banknote. This effect, of samples or exhibits of banknotes consisting of

bundles of banknotes with different levels of contamination, can also be seen in figure 5.18. The trace plot shows two sustained periods of high contamination, between the 30th and 50th banknotes, and between the 60th and 80th banknotes. Most of the remainder of the banknotes have log peak areas of less than 7, which, as can be seen from figure 5.15, would make them consistent with banknotes from general circulation.

The bottom plot in figure 5.17 shows the conditional density function of exhibit 7, which has 276 detected banknotes. This plot has just one main ridge, extending diagonally, which indicates that the contamination of this exhibit could be modelled using a more straightforward autoregressive process, with just one mean level, as for the sample associated with figure 5.16.

5.7 Conclusion

The two propositions being considered for a set of questioned banknotes **z**, that have been seized by a law enforcement agency, are:

- H_C: that the banknotes **z** are associated with a person who is associated with a drug crime involving cocaine, and
- H_B: that the banknotes **z** are associated with a person who is not associated with a drug crime involving cocaine.

Data have been collected to form two training data sets associated with these propositions: B, data associated with proposition H_B and C, data associated with proposition H_C. In this chapter, the steps taken to obtain these training data sets, B and C, were described. The data in both training sets form a convenience sample, and were obtained over many years. Samples in B are samples of banknotes from general circulation, such as from banks or shops. Exhibits in C were obtained from law enforcement agencies and are such that the suspect associated with the exhibit was later convicted of a crime involving cocaine. This definition of the exhibits in C differs from the definition used in previous analyses of these data. Previously (Besson (2004); Dixon et al. (2006)) banknotes seized by law enforcement agencies were used as the set C, without requiring that the suspect with whom the banknotes were associated was convicted of a crime involving cocaine. The main disadvantage of this previous approach is that the questioned sample, having been itself seized by law enforcement agencies, is by definition in the set C. Evaluating the likelihood ratio for the questioned sample will therefore add nothing to the analysis. There are 193 samples of banknotes in B, with each sample containing between 20 and 257 banknotes. There are 70 exhibits in C, with each exhibit containing between 20 and 1099 banknotes. These exhibits come from 29 different cases (crimes).

The data for each sample or exhibit take the form of the so-called log peak area for the cocaine product ion m/z 105, with one log peak area for each banknote within the sample or exhibit. The peak area is the sum of the number of gas phase ion transitions per second across a peak, where one peak corresponds to one banknote. Measurements of gas phase ion transitions were made using tandem

mass spectrometry. An algorithm was developed, based on the MassSpecWavelet algorithm of Du et al. (2006), which detects these peaks and calculates the associated peak areas.

Most banknotes in general circulation are contaminated with cocaine (Jourdan et al. (2013)) so methods are required for the evaluation of evidence that consider quantities of contamination rather than the proportion of contaminated banknotes within a sample or exhibit. It was shown in Section 5.6.1, that some of the crime exhibits are contaminated with similar quantities of cocaine as samples from general circulation. This is as a result of the definition of C. Banknotes that are associated with a person who was convicted of a crime involving cocaine might still be from general circulation. It was shown in Section 5.3.3 that the result of this overlap in contamination quantities between the two sets is that any model developed for discriminating between the two populations (B and C) will have high rates of misleading evidence for crime exhibits, even if the model fits the data well.

Measurements of cocaine on the banknotes were taken sequentially, with the order of the measurements taken generally corresponding to the order of the banknotes in the sample or exhibit. It has been shown that drugs can transfer between surfaces (Ebejer, Winn et al. (2007); Carter et al. (2003)) , suggesting that autocorrelation might be present within samples and exhibits. The sample partial autocorrelation function was calculated at a variety of lags for each sample and exhibit in Section 5.6.2. It was found that 77% of the crime exhibits and 89% of the general circulation samples had a significant sample partial autocorrelation coefficient at lag one (at the 5% significance level). This dropped to 14% of the crime exhibits and 24% of the general circulation samples at lag two. These results imply that the autoregressive process of lag one described in Chapter 3 could be used to model these data.

Samples and exhibits of banknotes often consist of multiple bundles of banknotes, each of which could have originated from a different location. Analysis of the data suggested that some samples and exhibits consist of banknotes which are contaminated at different levels; these different levels could relate to different levels of contamination on different bundles. Different sets of parameters will be needed to model these different levels of contamination. Because the different levels of contamination are thought to arise from the bundling together of banknotes (i.e. banknotes which are spatially close together) a standard mixture model is not suitable for the data. A model which accounts for dependence between the sets of parameters used for adjacent banknotes is required. A hidden Markov model, which has a latent state for each banknote, with the latent state determining the level of contamination ('low' or 'high') can be used. The model described in Section 3.13 is such a model; this model also allows for autocorrelation between adjacent banknotes.

The attributes of the data described in Section 5.6 were the motivation behind the development of the models described in Chapter 3. All of the models described use quantities of contamination rather than the proportion of banknotes contaminated. The autoregressive model and the nonparametric model take into account autocorrelation at lag one. The hidden Markov model also takes into account that samples and exhibits of banknotes often consist of multiple bundles, with different levels of contamination. Two levels of contamination are modelled, one corresponding to 'low' contamination

and one corresponding to ‘high’ contamination. In the following chapter, these models will be fitted to the data, and the results compared.

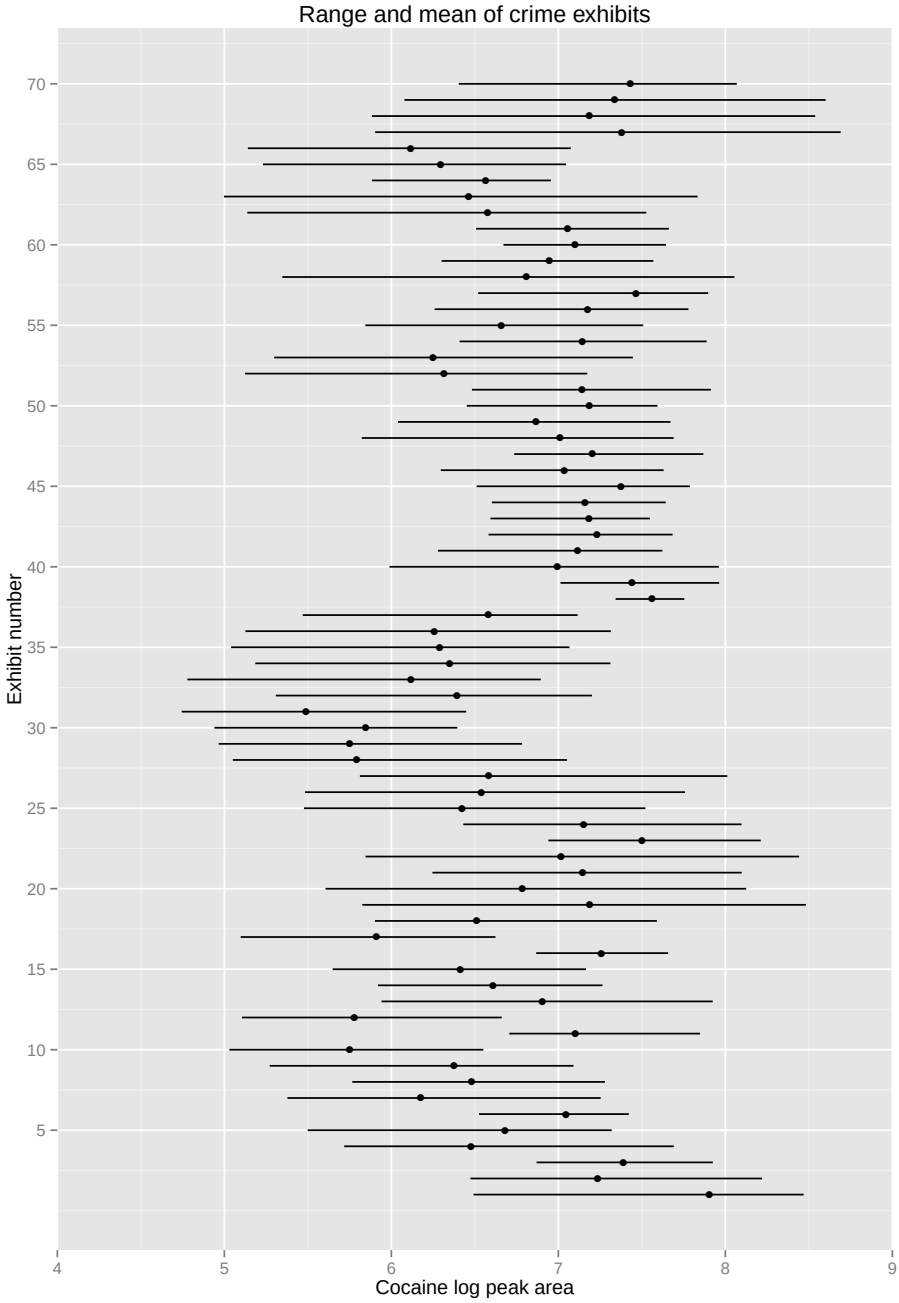

Figure 5.12: Plots of the ranges (line) and means (dot) of the 70 crime exhibits

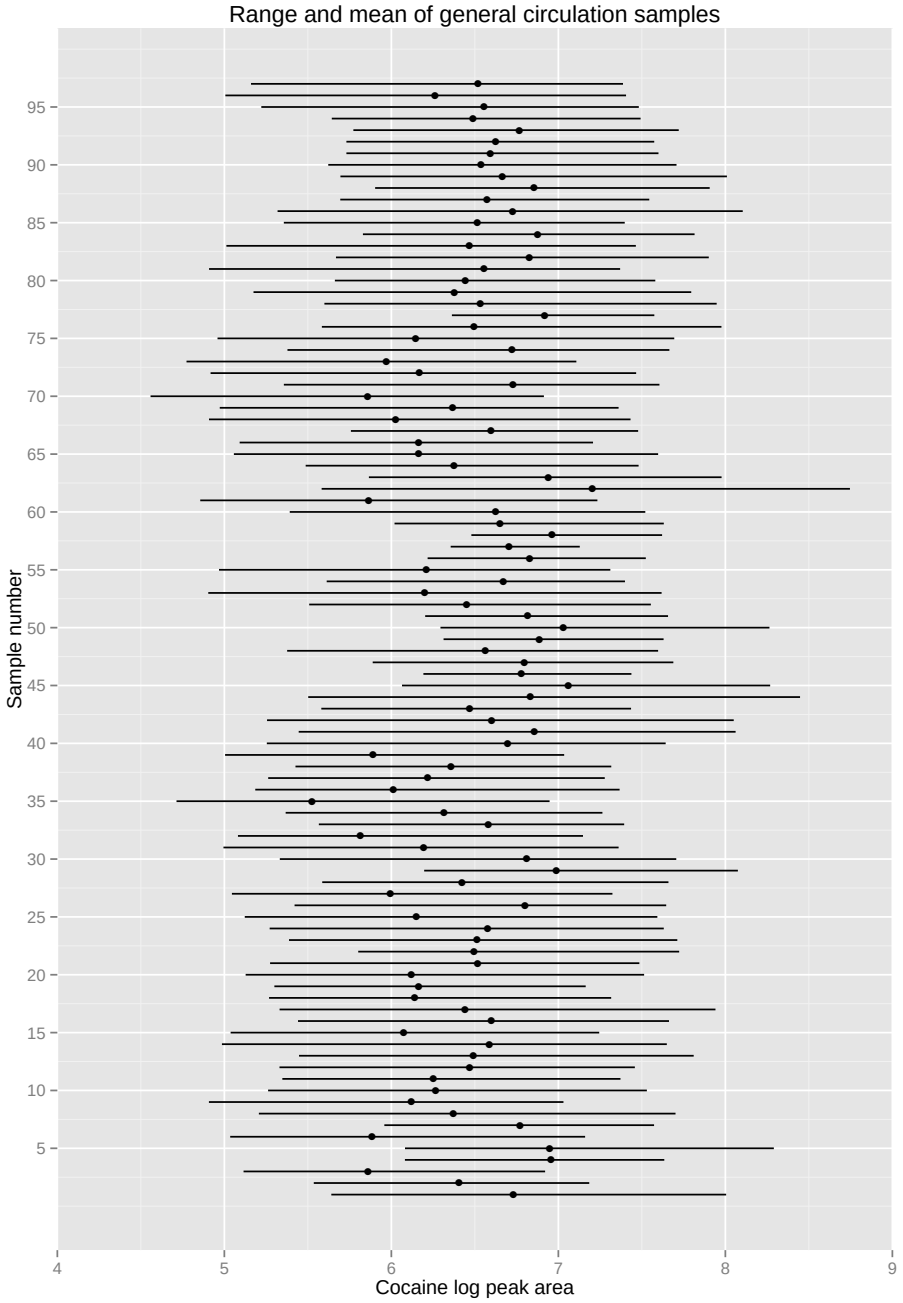

Figure 5.13: Plots of the ranges (line) and means (dot) of the first 97 general circulation samples

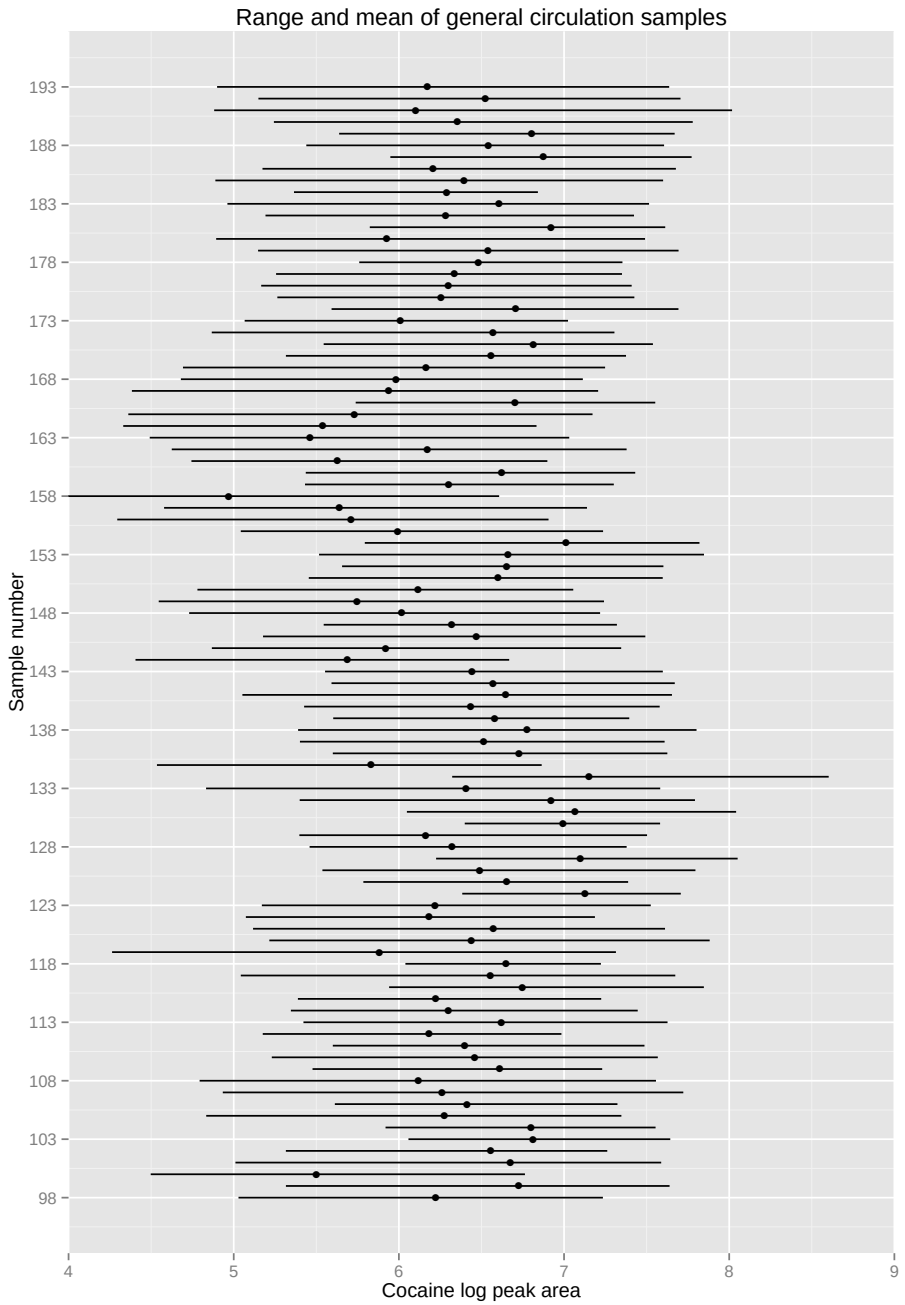

Figure 5.14: Plots of the ranges (line) and means (dot) of samples 98 to 193 from general circulation. Sample 158 has one banknote with a log peak area of 2.46; this outlier was not shown on the plot.

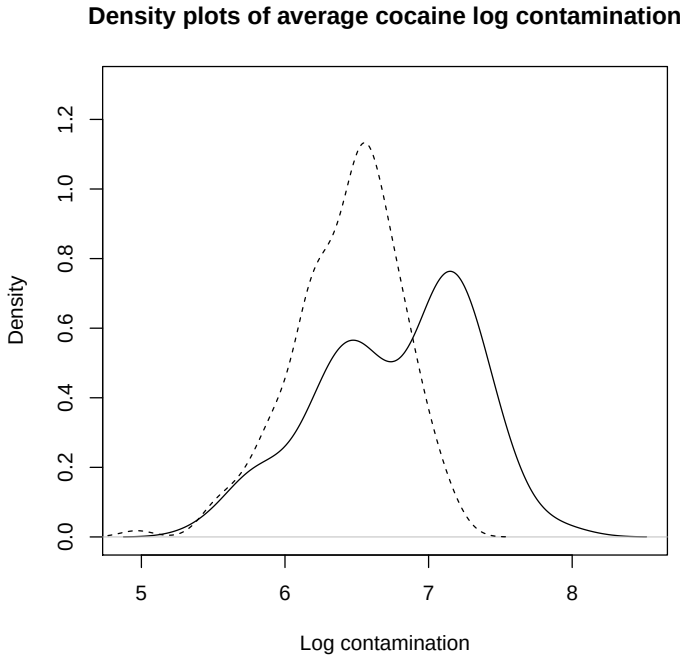

Figure 5.15: Density plots of mean contamination of samples/exhibits. Dashed line - general circulation, solid line - crime exhibit

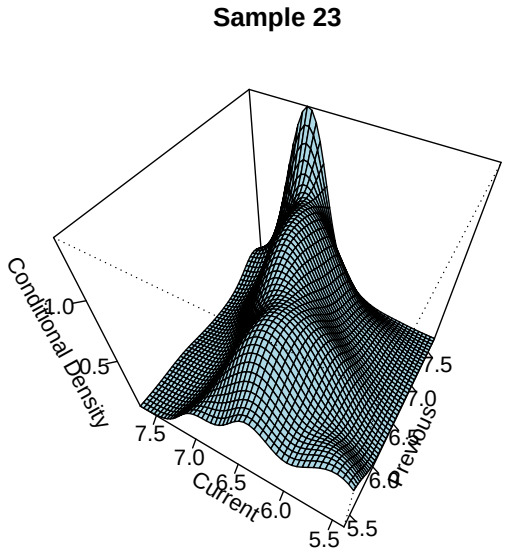

Figure 5.16: Conditional density plot of general circulation sample

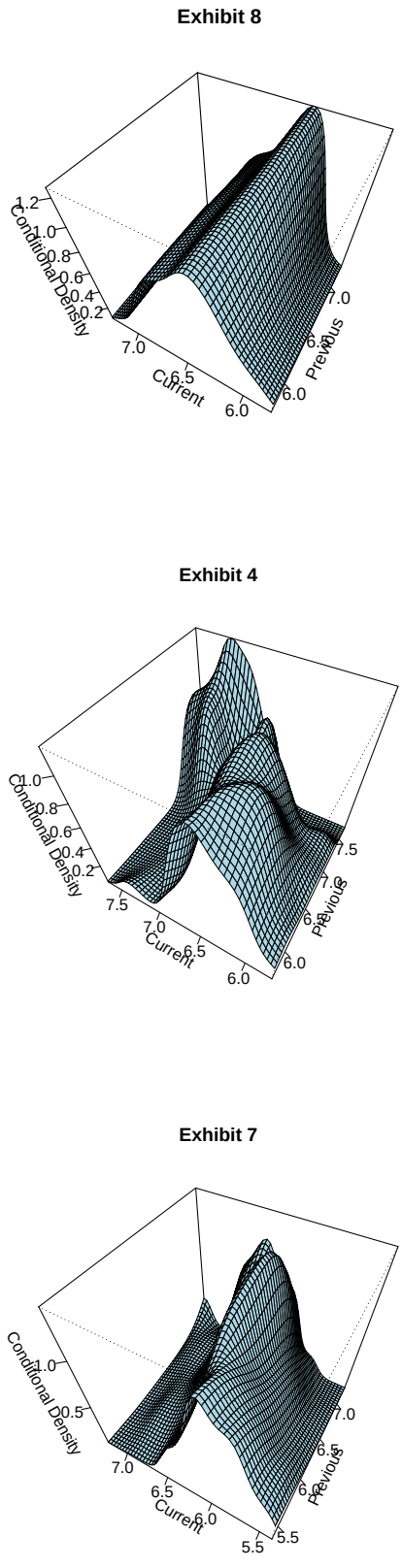

Figure 5.17: Conditional density plots of three crime exhibits

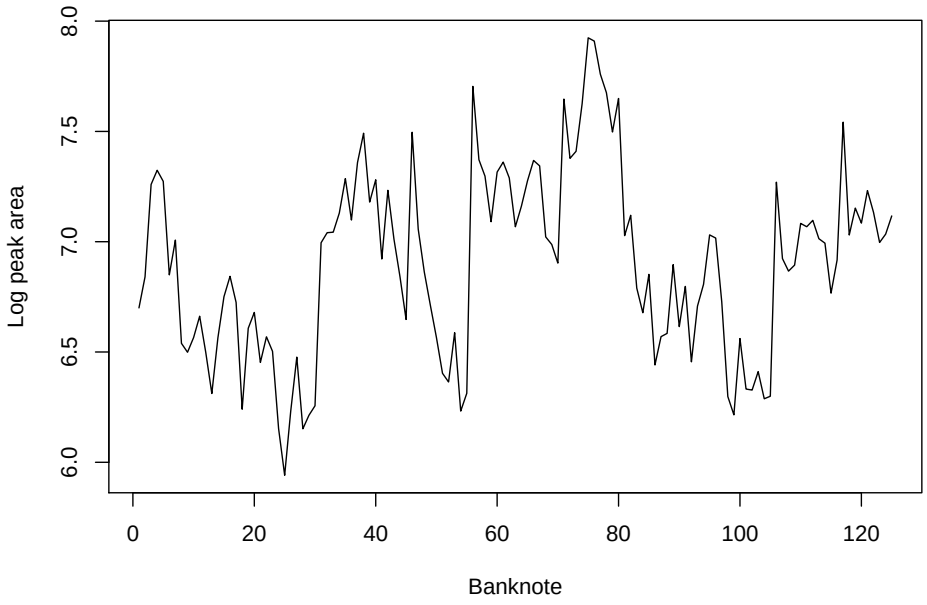

Figure 5.18: Trace plot of the log peak areas of exhibit 13

Chapter 6

Model fitting and evaluation

6.1 Introduction

In this chapter, the models detailed in Chapter 3 are fitted to the data introduced in Chapter 5, and the relative performances of these models are compared. The first model, an autoregressive process of order one, is fitted to all crime exhibits and all general circulation samples. Results are obtained both for the fixed effects model detailed in Section 3.1 and the random effects model detailed in Section 3.2. The second model, a hidden Markov model, is fitted to some exhibits and samples, with the autoregressive model being fitted to the remainder. The model selection is done using Bayes factors, as discussed in Section 4.4. A latent variable is used to describe whether each banknote has high or low levels of contamination; these latent variables are the hidden states of the hidden Markov model. The third model uses a nonparametric kernel density approach to estimate the conditional density function of the log peak area of a banknote, conditional on the log peak area of the previous banknote. Two different bandwidth selection methods are used: a fixed bandwidth and a variable bandwidth. A standard model, assuming independence between measurements on adjacent banknotes is also fitted to the data. Likelihood ratios are evaluated, treating each sample in training data set B and each exhibit in training data set C as the seized sample in turn, using the methods in Chapter 4 for each of the models.

One way of measuring the relative performance of the methods is by calculating rates of misleading evidence, which are determined by testing the methods on data from known populations (i.e. either known to be from B or C). Evidence can be misleading in one of two ways. A questioned sample known to have been associated with a person who is associated with crime involving cocaine can provide support for proposition H_B, that the banknotes are associated with a person who is not associated with crime involving cocaine. Alternatively, a questioned sample known to be associated with a person who is not associated with crime involving cocaine can provide support for proposition H_C. The rates of misleading evidence in this latter scenario, where samples of banknotes from general circulation provide support for proposition H_C, will be used as a measure of performance. Rates

of misleading evidence where crime exhibits provide support for proposition H_B will not be used as a measure of performance because of the known difficulty in achieving low rates of misleading evidence for crime exhibits, even for models which fit the data well, as described in Section 5.3.3. Other measures of performance include Tippett plots of the likelihood ratio values and scatter plots of the likelihoods of each proposition. Tippett plots consider the proportion of samples and the proportion of exhibits which have likelihood ratio values of greater than a given value. Tippett plots allow evaluation of the actual values of the likelihood ratios, rather than just the rates of misleading evidence. Scatter plots, where for each sample and exhibit the likelihood of H_C is plotted against the likelihood of H_B, will also be considered.

The notation used in this chapter is the same as the notation used in Chapters 3, 4 and 5. A crime is thought to have been committed. Part of the evidence is a sample of banknotes. The police suspect that cocaine is associated with this crime. The banknotes are analysed. The evidential data are given by the logarithms of the peak areas for the cocaine product ion *m/z* 105. One peak area is obtained for each of the banknotes (or on the subset of the banknotes that has been analysed). The likelihood ratio associated with the two propositions H_B and H_C (as given in Chapter 5) is to be calculated for this evidence.

Two sets of data, the training data sets, are used for development of the models; these were described in Chapter 5 and, analogously to the general data sets used in Chapter 3, are given by:

- Set B, consisting of data $\mathbf{x} = \{x_{it}; i = 1,\ldots,m_B, t = 1,\ldots,n_{B_i}\}$: the logarithms of the peak areas of cocaine product ion *m/z* 105 for banknotes from general circulation; there are m_B samples with n_{B_i} notes in sample i. Measurements on the i-th sample are denoted $\mathbf{x}_i$.
- Set C, consisting of data $\mathbf{y} = \{y_{it}; i = 1,\ldots,m_C, t = 1,\ldots,n_{C_i}\}$: the logarithms of the peak areas of cocaine product ion *m/z* 105 for banknotes from crime exhibits; there are m_C exhibits with n_{C_i} notes in exhibit i. Measurements on the i-th exhibit are denoted $\mathbf{y}_i$.

The questioned (or seized) evidential sample is

- $\mathbf{z} = (z_1, z_2, \ldots, z_n)$: the logarithms of the peak areas of cocaine product ion *m/z* 105 for a sample of n banknotes seized by law enforcement agencies, to form one exhibit.

6.2 Fitting models to the training data

6.2.1 Autoregressive model

In Section 5.6.2 it was shown that the majority of samples in set B and exhibits in set C were autocorrelated, and that an autoregressive model of order one might be appropriate to model the data in each of these sets. The method for fitting an autoregressive model of order one to a single sample was given in Section 3.1. In this section, details specfic to the fitting of this model to data relating to traces of cocaine on banknotes will be given. The model was fitted to each sample $\mathbf{x}_i$ for $i \in \{1,\ldots,m_B\}$ and

each exhibit $\mathbf{y}_i$ for $i \in \{1,\ldots,m_C\}$ to obtain draws from the posterior density functions $f(\theta^B_{A_i} \mid \mathbf{x}_i)$ and $f(\theta^C_{A_i} \mid \mathbf{y}_i)$ for each i, where $\theta^B_{A_i} = (\mu^B_i, (\sigma^B_i)^2, \alpha^B_i, \beta^B_i)$ and $\theta^C_{A_i} = (\mu^C_i, (\sigma^C_i)^2, \alpha^C_i, \beta^C_i)$.

Prior distributions

The prior distributions used for the general autoregressive model are given in Section 3.1.1. For a general sample or exhibit of banknotes $\mathbf{w}_i$, the specific priors that were used are given by:

- $\mu \sim \mathrm{N}(\frac{1}{2}(\max(\mathbf{w}_i) + \min(\mathbf{w}_i)), \mathrm{range}(\mathbf{w}_i)^2)$;
- $\sigma^2 \sim \mathrm{IG}(2.5, \beta)$, where IG denotes the inverse gamma distribution and β is as a hyperparameter;
- $\beta \sim \Gamma(0.5, 4/\mathrm{range}(\mathbf{w}_i)^2)$;
- $\alpha \sim \mathrm{N}(0, 0.25)$, with the autocorrelation restricted to lie between -1 and 1.

These prior distributions are used to provide compatibility with the hidden Markov model. The prior distributions and hyperparameters associated with σ^2, μ and β have the same form as those in Richardson and Green (1997) and Rydén (2008). The hyperparameters for μ and β are dependent on the data $\mathbf{w}_i$. Ideally, these hyperparameters would instead be chosen by an expert and would not be dependent on the data. In Richardson and Green (1997), the choices of the priors and hyperparameters are described as 'making only 'minimal' assumptions on the data', a conclusion which is also true here. The prior distribution of the parameter μ is relatively uninformative over the range of sensible values for the mean of the log peak areas; none of the samples or exhibits has a mean which is outside the range $(4.9, 8.0)$ and the range of each sample and exhibit is generally around two. Similarly, since the range of the data $\mathbf{w}_i$ is around two, the prior distribution of β has most of its weight in the range $(0, 2)$. Values of β smaller than two result in prior distributions for σ^2 which do not place a large weight on values of σ^2 that are larger than around three. This is a reasonable assumption, which is not very informative, given that all of the banknotes in both data sets have log peak areas that lie in the range $(2, 9)$.

Metropolis-Hastings sampler

Methods for obtaining draws from the posterior density functions $f(\theta^B_{A_i} \mid \mathbf{x}_i)$ for $i \in \{1,\ldots,m_B\}$ and $f(\theta^C_{A_i} \mid \mathbf{y}_i)$ for $i \in \{1,\ldots,m_C\}$ are given in Section 3.1.2. For the training data sets B and C, the Gibbs sampler did not perform well for the hidden Markov model, tending to become trapped in local modes for small samples. For consistency, the Metropolis-Hastings sampler was therefore used for both the autoregressive model and the hidden Markov model.

The range of variances, V_k used for the multivariate Normal proposal distribution (see Section 3.1.2) are given by:

Parameter	Range of V_k values	
	Crime exhibits	Background samples
μ	(0.0495, 0.15)	(0.03, 0.2)
σ^2	(0.15, 0.35)	(0.1, 0.2)
α	(0.1, 0.35)	(0.12, 0.2)
β	(0.45, 0.675)	(0.3, 0.5)

These values were obtained by manual adjustment, so that acceptance probabilities for all samples and exhibits in initial runs of the sampler were approximately 25%.

In total, 750,000 iterations of the sampler were carried out for each sample and exhibit. The first 250,000 of these 750,000 iterations were discarded as burn-in. A thinning parameter of 25 was used to remove autocorrelation from within the chain. This left 20,000 draws from $f(\theta \mid \mathbf{w}_i)$ (where θ is equal to $\theta^B_{A_i}$ or $\theta^C_{A_i}$, depending on whether data $\mathbf{x}_i$ or $\mathbf{y}_i$ are being considered). Five chains of this length were run from different starting positions; these starting positions were sampled from the prior distributions. The upper confidence limits for the univariate Gelman-Rubin statistic for each parameter (Gelman and Rubin (1992)) and the multivariate extension of the Gelman-Rubin statistic (Brooks and Gelman (1998)) were calculated for each sample and each exhibit using the Coda R package (Plummer et al. (2006)). The univariate Gelman-Rubin statistic monitors the ratio of estimates of the pooled-chain variance and the within-chain variance. The pooled chain variance estimate is a weighted average of the within-chain and the between-chain variances. To use this diagnostic tool for convergence, the starting points of the chains should be overdispersed in relation to the posterior distribution of the parameters. If these starting positions are overdispersed then the estimate of the pooled-chain variance will be an overestimate if the chain has not converged, as the between-chain variance will be large. As the chains converge, the ratio of the two variance estimates should tend to one from above. The multivariate Gelman-Rubin statistic extends this idea to covariance matrices by considering a measure of the difference between the within chain covariance matrix and an estimate of the pooled-chain covariance matrix. This measure also tends to one from above and is given in Brooks and Gelman (1998) by

$$\frac{N-1}{N} + \left(\frac{M+1}{M}\right)\lambda_1$$

where λ_1 is the largest eigenvalue of the matrix $W^{-1}B$, B is the between-chain covariance matrix, W is the within chain covariance matrix, N is the number of draws in a chain and M is the number of chains.

To use the Gelman-Rubin statistics, the marginal posterior distributions of the parameters should be approximately Normal. To achieve approximate Normality, logarithms were taken of the draws from the posterior distributions of the parameters σ^2 and β. All samples and exhibits had multivariate Gelman-Rubin statistics which were less than 1.0047, and univariate Gelman-Rubin statistics which

were well under the value of 1.1 recommended in Brooks and Gelman (1998). Trace plots of the chains were also visually inspected for convergence. On visual inspection, one general circulation sample (sample 172) was found not to have converged, which was also the sample with the largest multivariate Gelman-Rubin statistic of 1.0047. The posterior draws from this sample were removed from the analysis and not used to calculate likelihood ratios.

6.2.2 Autoregressive model with random effects

The autoregressive model with random effects, described in Section 3.2, was fitted to the training data sets B and C, for comparison with the approach taken in Section 6.2.1. For this model, draws from the posterior distribution of the parameter $\theta^B_{A_r}$, conditional on the entire set of training data $\mathbf{x}$ and draws from the posterior distribution of the parameter $\theta^C_{A_r}$, conditional on the entire set of training data $\mathbf{y}$ were obtained.

Prior distributions

The prior distributions used for the general autoregressive model with random effects are given in Section 3.2.1. The specific prior distributions that were used for this model are given by:

- $\mu_\mu \sim \mathrm{N}(6, 20)$
- $\sigma_\mu \sim \mathrm{U}(0, 5)$
- $\gamma_V \sim \Gamma(0.1, 0.1)$
- $\beta_V \sim \Gamma(0.1, 0.1)$
- $\mu_\alpha \sim \mathrm{N}(0, 4)$
- $\sigma_\alpha \sim \mathrm{U}(0, 5)$.

These prior distributions were chosen based on advice in Gelman (2006), which suggests the use of a uniform prior for the hierarchical standard deviation in a random effects model. The hyperparameters used for these prior distributions are relatively uninformative; different parameters were tested, and it was found that the parameters chosen had little effect on the posterior distributions obtained.

Sampling from the posterior distributions

The R package rjags (Plummer (2013)) was used to obtain draws from the posterior density functions $f(\theta^B_{A_r} \mid \mathbf{x})$ and $f(\theta^C_{A_r} \mid \mathbf{y})$. The code used to obtain these draws is provided in Appendix C. The number of draws obtained from the sampler depended on the number of draws required to get a good estimate of the likelihood ratio (this is discussed further in Section 6.3.2), but 10,000 draws were discarded as burn-in each time the sampler was run.

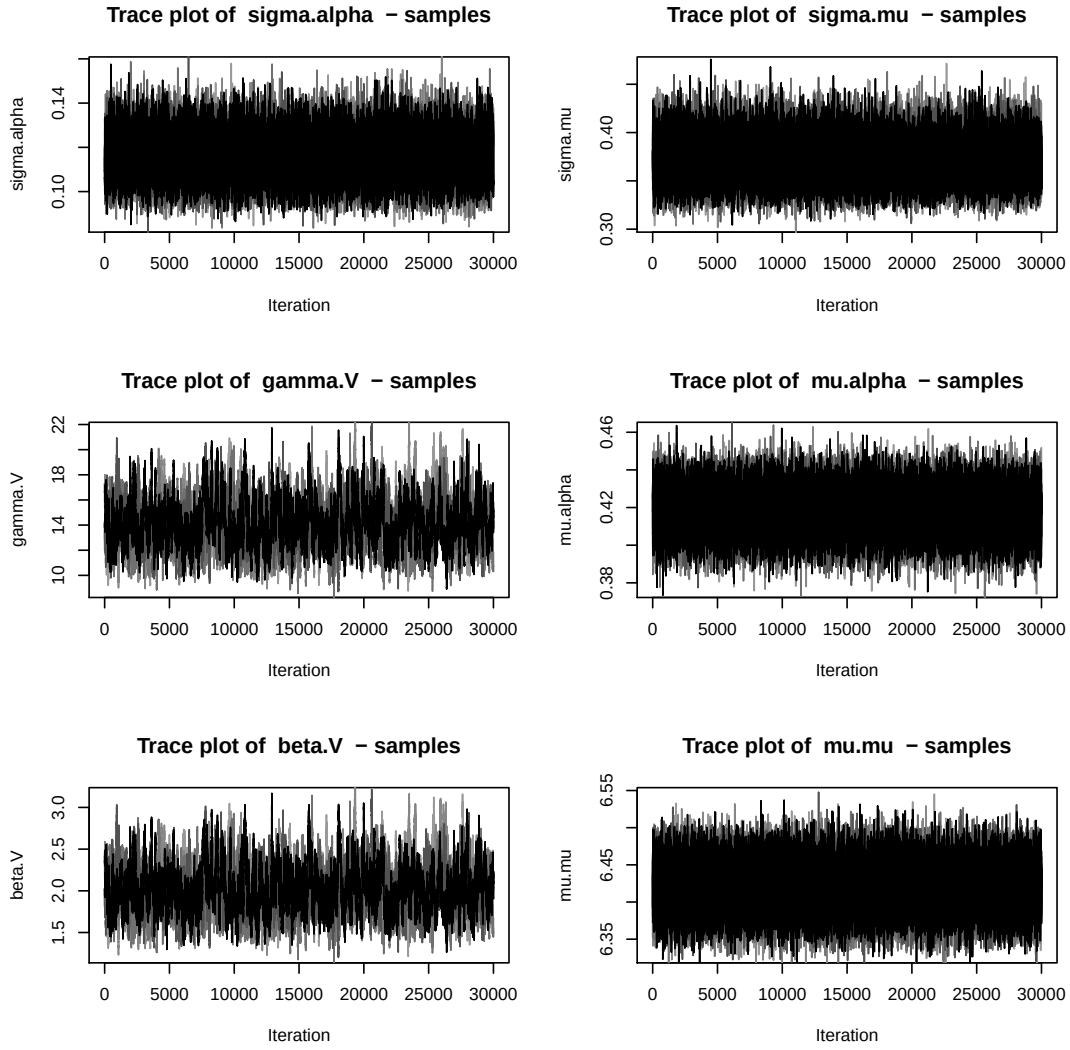

Figure 6.1: Trace plots showing five chains from sampler drawing from posterior distribution of parameters in $\theta^B_{A_r}$ (different chains are shown in different shades); general circulation parameters for autoregressive model with random effects

To check convergence, five chains of length 30,000 were obtained from the sampler for each of the posterior density functions $f(\theta^B_{A_r} \mid \mathbf{x})$ and $f(\theta^C_{A_r} \mid \mathbf{y})$. Univariate and multivariate Gelman-Rubin statistics were calculated; the largest of these was found to be 1.01, which is much smaller than the suggested value of 1.1. Trace plots and density plots of each of the parameters in $\theta^B_{A_r}$ and $\theta^C_{A_r}$ were analysed for convergence. The trace plots for the parameters in $\theta^B_{A_r}$ can be seen in figure 6.1. Trace plots for parameters in $\theta^C_{A_r}$ can be seen in figure 6.2. These plots show the five chains before any draws have been discarded for burn-in, and show that the chains have converged for all parameters after 10,000 iterations. Density plots of the marginal posterior density functions of each of the parameters in $\theta^B_{A_r}$ and $\theta^C_{A_r}$ can be seen in figures 6.3 and 6.4. The similarity of the marginal posterior density functions for each of the chains implies convergence for all parameters.

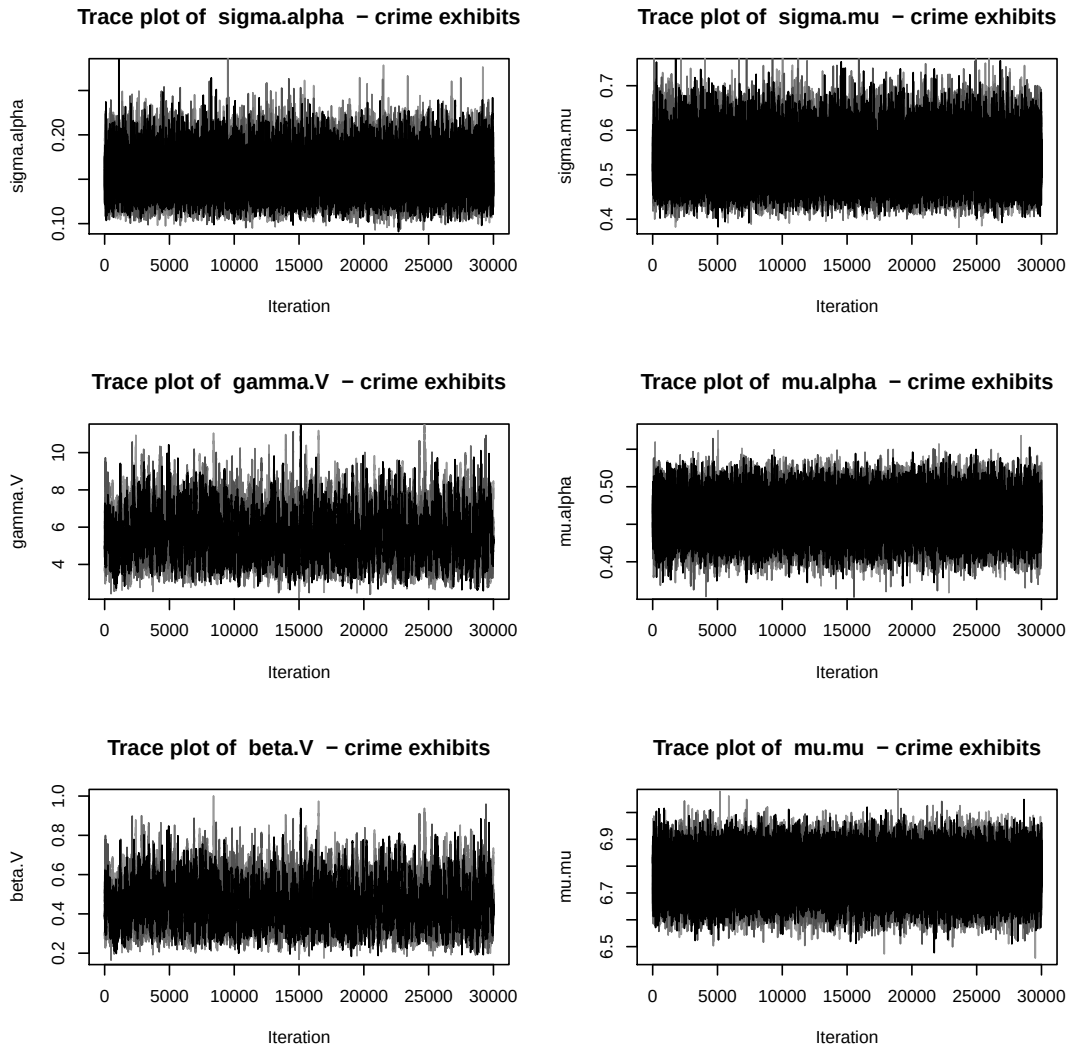

Figure 6.2: Trace plots showing five chains from sampler drawing from posterior distribution of parameters in $\theta^{C}_{A_r}$ (different chains are shown in different shades); crime parameters for autoregressive model with random effects

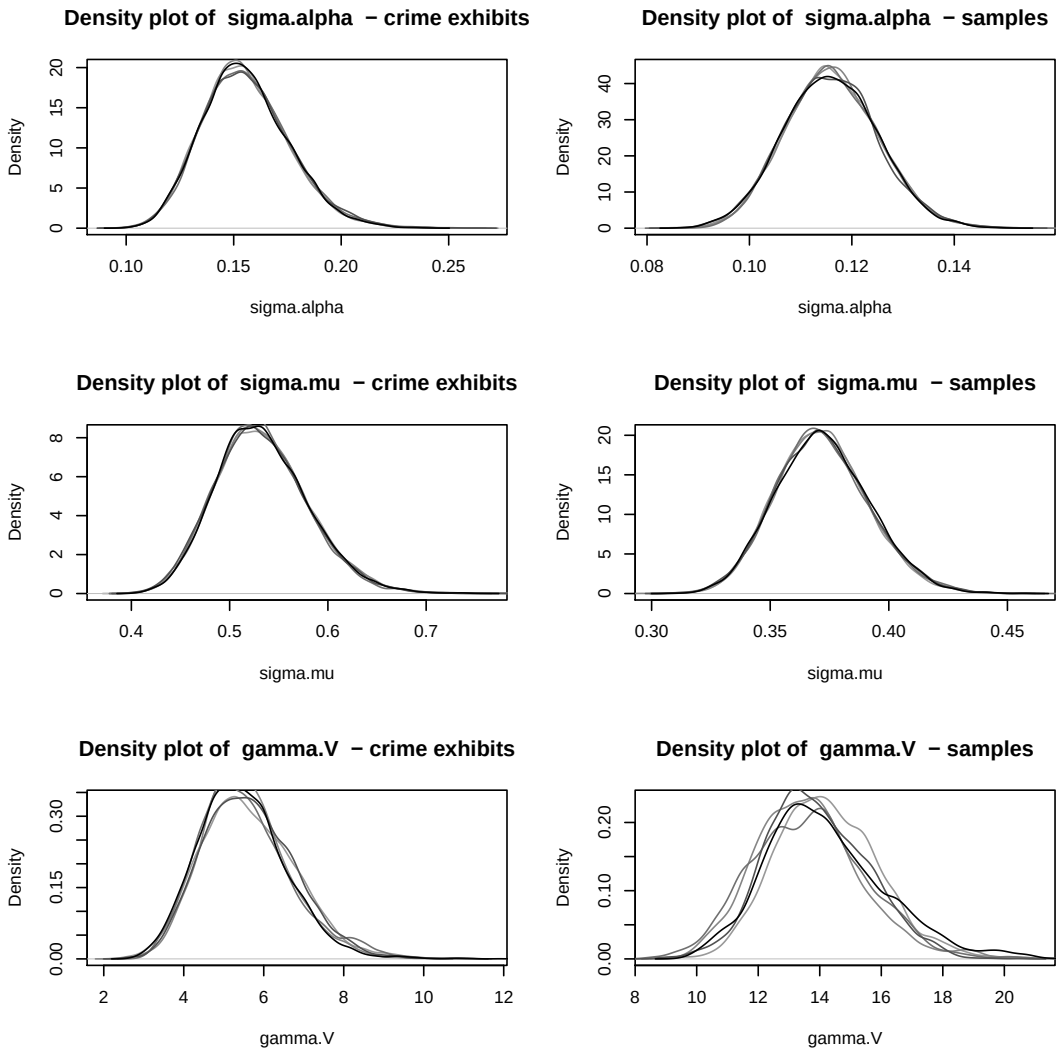

Figure 6.3: Density plots estimated from each of five chains from the sampler drawing from posterior distributions of parameters in $\theta^B_{A_r}$ and $\theta^C_{A_r}$ for autoregressive model with random effects

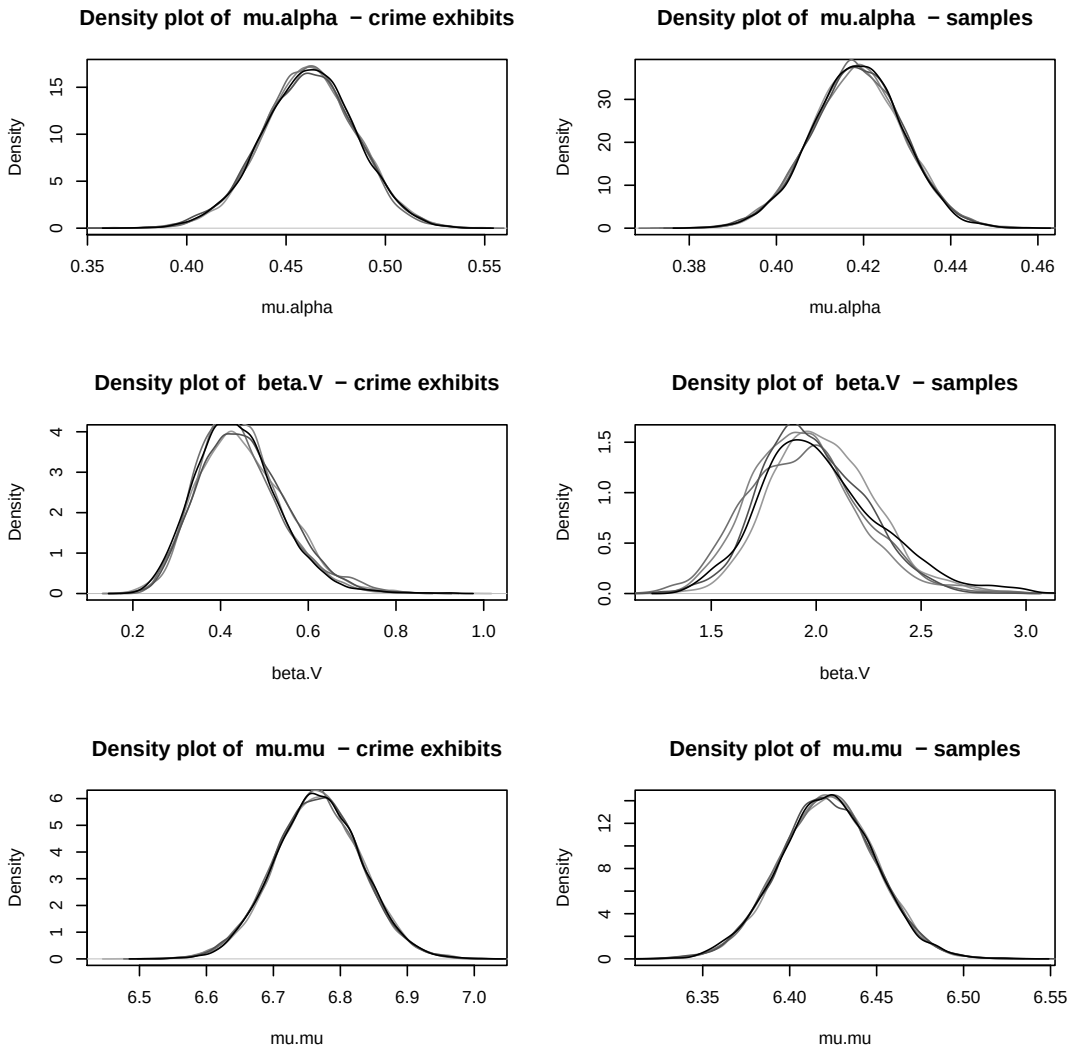

Figure 6.4: Density plots estimated from each of five chains from the sampler drawing from posterior distributions of parameters in $\theta^B_{A_r}$ and $\theta^C_{A_r}$ for autoregressive model with random effects

6.2.3 Hidden Markov model

In Section 5.6.3 it was shown that a hidden Markov model can be used to model the different contamination levels on different bundles of banknotes within samples and exhibits. The method for fitting a hidden Markov model that is able to model two different contamination levels, in addition to autocorrelation, to a general sample $\mathbf{w}_i$ is described in Section 3.3. For data relating to traces of cocaine on banknotes, the observations are given by the logarithms of the peak areas relating to the cocaine product ion m/z 105 for each banknote. A latent state is associated with each banknote. The states are used to model the contamination level of each banknote (either high or low) within a sample or exhibit. A mean and variance parameter is associated with each contamination level. Returning to the notation used in Section 3.3, μ_1 and σ_1^2 denote the mean and variance parameters associated with a low contamination level, and μ_2 and σ_2^2 denote the mean and variance parameters associated with a high contamination level. The latent states form a Markov chain, with an associated transition matrix, which allows for the modelling of the clustering together of banknotes in bundles. It is thought that banknotes in the same bundle are likely to have the same state. The model in Section 3.3 was fitted to each sample $\mathbf{x}_i$ for $i \in \{1,\ldots,m_B\}$ and each exhibit $\mathbf{y}_i$ for $i \in \{1,\ldots,m_C\}$ to obtain estimates of the posterior density functions $f(\theta^B_{H_i} \mid \mathbf{x}_i)$ and $f(\theta^C_{H_i} \mid \mathbf{y}_i)$ for each i.

Prior distributions

The prior distributions used for the general hidden Markov model are given in Section 3.3.4. For a general sample or exhibit with n_{D_i} banknotes, $\mathbf{w}_i$, the specific prior distributions used for data relating to traces of cocaine on banknotes are given by:

- $\mu_1, \mu_2 \sim \mathrm{N}\left((\max(\mathbf{w}_i) + \min(\mathbf{w}_i))/2, \mathrm{range}(\mathbf{w}_i)^2\right)$,
- $\sigma_1^2 \sim \mathrm{IG}(2.5, \beta_1)$,
- $\sigma_2^2 \sim \mathrm{IG}(2.5, \beta_2)$,
- $\beta_1, \beta_2 \sim \Gamma\left(0.5, 4/\mathrm{range}(\mathbf{w}_i)^2\right)$.
- $\alpha \sim \mathrm{N}(0, 0.25)$, with the autocorrelation restricted to lie between -1 and 1.
- $p_{01}, p_{10} \sim \mathrm{Beta}(0.6, 4)$, and restricted to lie between $2/n_{D_i}$ and $(n_{D_i} - 2)/n_{D_i}$.

The hyperparameters in the prior distribution for p_{01} and p_{10} were chosen following the advice in Spezia (2010). Transition probabilities were thought to be small because of the clustering together of banknotes in the same bundle; a beta prior distribution with parameters 0.6 and 4 assigns most weight to transition probabilities smaller than 0.5. A discussion of the choice of a truncated beta prior distribution for p_{01} and p_{10} is given in Section 3.3.4. The hyperparameters used in the prior distributions for the means and variances are similar to those used in Richardson and Green (1997) and Rydén (2008). The dependence of some of the prior distributions on the data $\mathbf{w}_i$ is discussed in

Section 6.2.1. The exact hyperparameters used were chosen after the prior distributions were tested on simulated data with similar properties to the real data.

Metropolis-Hastings sampler

Methods for obtaining draws from the posterior density functions $f(\theta^B_{H_i} \mid \mathbf{x}_i)$ for $i \in \{1,\ldots,m_B\}$ and $f(\theta^C_{H_i} \mid \mathbf{y}_i)$ for $i \in \{1,\ldots,m_C\}$ are given in Section 3.3.6. As discussed in Section 6.2.1, the Gibbs sampler did not perform well with the data sets relating to cocaine traces on banknotes so the Metropolis-Hastings sampler was used to obtain these draws.

The range of variances, V_k, used for the multivariate Normal proposal distribution (see Section 3.3.6) are given by:

Parameter	Range of V_k values	
	Crime exhibits	Background samples
μ	(0.033, 0.1)	(0.03, 0.133)
σ^2	(0.1, 0.233)	(0.1, 0.2)
α	(0.067, 0.233)	(0.08, 0.133)
p	(0.25, 0.5)	0.5
β	(0.3, 0.45)	(0.2, 0.333)

Five chains, each consisting of 20,000 draws from $f(\theta \mid \mathbf{w}_i)$ were obtained, with burn-in and thinning parameters as in section 6.2.1. The univariate and multivariate Gelman-Rubin statistics (Gelman and Rubin (1992); Brooks and Gelman (1998)), as also described in Section 6.2.1, were calculated to check convergence. For the hidden Markov model, posterior density functions of all parameters with the exception of α should be bimodal (see Section 3.3.6 for more details), which is a violation of the Normality assumption required for use with the Gelman-Rubin statistic. As such, trace plots of the chains and marginal posterior density plots of the parameters were also visually inspected for convergence. The draws from any sample or exhibit which did not converge (either as judged from visual inspection, or those with a Gelman-Rubin statistic greater than 1.1) were removed from the analysis and not used to calculate likelihood ratios.

As discussed in Section 3.3.4, convergence was difficult to achieve for some samples and exhibits; it was suspected that this was due to the overfitting of the model (with too many states). To resolve these problems, methods were given in Section 4.4 that allowed some of the samples and exhibits to be fitted with a hidden Markov model, and some to be fitted with an autoregressive model. Model selection was made using Bayes' factors and is discussed further in Section 6.3.3. In total, 40 crime exhibits of a total of 70, and 94 samples of a total of 193 from general circulation were fitted with a hidden Markov model. An autoregressive model was used for the remainder. Of those samples and exhibits modelled with the hidden Markov model, three exhibits of the 40 (exhibits 6, 45 and 50) and five samples of the 94 (samples 44, 60, 124, 134 and 158) were removed from the analysis due to

nonconvergence (as judged by visual inspection or by the Gelman-Rubin statistic). Of those samples and exhibits modelled with an autoregressive model, none was removed due to nonconvergence. Sample 172, removed when all samples were modelled with an autoregressive model, was found to have a better fit with the hidden Markov model. Considering the Gelman-Rubin statistic alone, three exhibits in total (of all 70 exhibits) and 16 samples in total (of all 193 samples) had a Gelman-Rubin statistic of greater than 1.1 when draws were obtained for the hidden Markov model. Of these, two of the three exhibits and 15 of the 16 samples were found to have a better fit with the autoregressive model, i.e. the marginal likelihood of the autoregressive model was bigger than the marginal likelihood of the hidden Markov model. This supports the suggestion that nonconvergence of the samplers for some samples and exhibits (for the hidden Markov model) was being caused by overspecification of the model for those samples and exhibits.

6.2.4 Nonparametric model

The method for fitting a nonparametric model, which allows for lag one autocorrelation, to a general sample $\mathbf{w}_i$ is given in Section 3.4. This method dispenses with the assumption of Normality of errors (used in both the hidden Markov model and the autoregressive model). Estimates of conditional density functions, $\hat{f}_{D_i}(\cdot \mid \cdot)$, and marginal density functions, $\hat{f}_{D_i}(\cdot)$, are obtained for each sample in B and each exhibit in C.

As can be seen in the plots in figures 5.12, 5.13 and 5.14, quantities of contamination vary considerably between different samples and exhibits of banknotes. To calculate the likelihood ratio, each of the functions $\hat{f}_{D_i}(\cdot \mid \cdot)$ must be evaluated for each pair of banknotes in the questioned sample. As quantities of contamination can vary considerably, these functions may have to be evaluated at points that are very dissimilar to the data used to estimate the function, i.e. at points that are in the tails of the estimated conditional density function. As discussed in Section 3.4, better results may therefore be obtained by using a bandwidth which varies, depending on the amount of data present. Two bandwidth selection methods were described in Section 3.4, a fixed bandwidth and an adaptive nearest neighbour bandwidth. Results from these two methods of bandwidth selection will be compared.

6.3 Obtaining likelihood ratios

In this section the methods introduced in Chapter 4 for the evaluation of likelihood ratios for the autoregressive, hidden Markov, nonparametric and standard models will be applied to evidence relating to traces of cocaine on banknotes.

6.3.1 Autoregressive model

A method for the evaluation of the likelihood ratio when the autoregressive model is the assumed model for the data is given in Section 4.1. Draws $\theta_{A_i}^{C(r)}$ for $i \in \{1,\ldots,m_C\}$ and $r \in \{1,\ldots,N\}$ and $\theta_{A_i}^{B(r)}$ for $i \in \{1,\ldots,m_B\}$ and $r \in \{1,\ldots,N\}$ were obtained using the Metropolis-Hastings sampler described in 6.2.1. Each general circulation sample in B and each crime exhibit in C was treated as the seized sample $\mathbf{z}$ in turn. The draws associated with the remainder of the samples and exhibits (also removing draws associated with other exhibits in the same case if the seized sample was from set C, and draws associated with sample 172, the sampler of which did not converge) were then used to evaluate the likelihood ratio for this seized sample for the two propositions, H_C and H_B. As a result, the test sample $\mathbf{z}$ was always independent of the training samples B and C. The likelihood ratio for $\mathbf{z}$ was estimated using (4.1), with the individual integrals estimated using (4.3). The weights used were the suggested weights in Section 4.1 of $v_i = n_{D_i} / \sum_{i=1}^{m_D} n_{D_i}$.

Each integral in (4.1) was estimated with a randomly sampled (without replacement) set of 5,000 draws from the 20,000 posterior draws in the first chain of the Metropolis-Hastings sampler for each sample or exhibit. This process was repeated three times, again randomly sampling 5,000 draws without replacement from the 20,000 draws in the first chain (having replaced those draws used for previous estimates). Three different estimates of the likelihood ratio $f(\mathbf{z} \mid H_C)/f(\mathbf{z} \mid H_B)$ were then obtained from these integral estimates. The maximum and minimum of these three likelihood ratio evaluations were compared. If a seized sample had a ratio of the maximum estimate to the minimum estimate of greater than two, then the likelihood ratio was re-calculated, this time with the estimate of each integral in (4.1) based on all 20,000 posterior draws from the first chain. This repeat calculation was required for three of 263 seized samples (where each seized sample is a sample or exhibit from B or C). For these three seized samples, this process was repeated, obtaining likelihood ratio estimates for each of the five chains (so five likelihood ratios were obtained, each based on 20,000 posterior draws for each integral). The range of these values was calculated; the minimum and maximum likelihood ratios obtained for these three seized samples are presented in table 6.1. In practice, the likelihood ratio for a seized sample could be calculated from the full set of 100,000 posterior draws. A subset of these draws were used here to save computational time, as likelihood ratio estimates were required for 263 seized samples. The ranges in table 6.1 give an idea of the largest errors that would be found in the estimation of the likelihood ratio using 20,000 posterior draws. These ranges can be calculated in practice and used to check that a sufficient number of draws has been used for the estimation of the likelihood ratio. The likelihood ratios used to present the results are the average of the likelihood ratios obtained (so based on either 15,000 or 100,000 evaluations for each integral).

Methods discussed in Section 4.1 for the estimation of confidence intervals for the estimate of the numerator and denominator of the likelihood ratio were found not to be suitable here. Evaluations of the N values of $f^{(r)}(\mathbf{z} \mid H_D)$ (see Section 4.1 for definition) were not approximately Normally distributed for some samples and exhibits, and so confidence intervals based on an assumption of Normality

Sample / exhibit	Min LR obtained	Max LR obtained
39 (exhibit)	2.37×10^4	8.99×10^4
67 (exhibit)	1.37	2.30
158 (sample)	3.89	1.55×10^2

Table 6.1: Range of likelihood ratio values obtained for samples or exhibits requiring a repeat calculation of the likelihood ratio when treated as the seized sample (AR(1) model).

Sample / exhibit	Min LR obtained	Max LR obtained
39 (exhibit)	3.22×10^6	1.03×10^9
67 (exhibit)	1.92×10^2	6.87×10^3
69 (exhibit)	2.10×10^2	1.06×10^3
158 (sample)	8.70×10^{-1}	9.04

Table 6.2: Range of likelihood ratio values obtained for samples or exhibits with large variation in the estimate of the likelihood ratio after 500,000 posterior draws (AR(1) model with random effects).

were not appropriate. Using multiple estimates of the function $f(\mathbf{z} \mid H_D)$ and forming confidence intervals based on quantiles was not practical due to long computation times, especially because a confidence interval would be required for all 263 samples and exhibits.

6.3.2 Autoregressive model with random effects

A method for the evaluation of the likelihood ratio when the autoregressive model with random effects is the assumed model for the seized sample was given in Section 4.2. As before, each general circulation sample in B and each crime exhibit in C was treated as the seized sample $\mathbf{z}$ in turn. The remainder of the samples and exhibits (those not in the same case as the seized sample) were then used as the training data sets B and C, to obtain draws $\theta_{A_r}^{B(r)}$ and $\theta_{A_r}^{C(r)}$ for $r \in \{1,\ldots,N\}$ using the Metropolis-Hastings sampler described in Section 6.2.2. The integrals (4.4) for H_C and H_B were then estimated using these draws. The number of draws, N, used to estimate the likelihood ratio for $\mathbf{z}$ varied from 20,000 to 500,000. For a given N, the process described in Section 6.3.1 was used to determine whether more draws were required to obtain a better estimate of the likelihood ratio. For the random effects model, a new sampler had to be run for each seized sample (as removing the seized sample from the data set with which it is associated changes this data set) so there was greater scope to vary the value of N between samples. For the standard autoregressive model, the sampler was run just once, so the number of draws was fixed at 20,000 per chain before estimation of the likelihood ratio for all of the questioned samples.

After 500,000 draws, some of the samples and exhibits, when treated as seized samples, still had large variation in their likelihood ratio estimates. The ranges found for the likelihood ratios for these samples and exhibits are shown in table 6.2. These ranges are based on five estimates of the likelihood ratio, so that comparisons can be made with the other models used. As described in Section 4.2, more draws are required to estimate likelihood ratios for the autoregressive model with random effects in comparison to the autoregressive model without random effects. This is a reflection of the increased dimension of the integral to be estimated.

6.3.3 Using a combination of the hidden Markov model and the autoregressive model

A method for the evaluation of the likelihood ratio when some samples and exhibits in the training data sets are modelled with a hidden Markov model and the remainder are modelled with an autoregressive model is given in Section 4.4. Draws $\theta_{A_i}^{C(r)}$ and $\theta_{H_i}^{C(r)}$ for $i \in \{1,\ldots,m_C\}$ and $r \in \{1,\ldots,N\}$ and $\theta_{A_i}^{B(r)}$ and $\theta_{H_i}^{B(r)}$ for $i \in \{1,\ldots,m_B\}$ and $r \in \{1,\ldots,N\}$ were obtained using the Metropolis-Hastings samplers described in Sections 6.2.1 and 6.2.3. Bayes factors were calculated to decide whether each sample and exhibit was to be modelled using a hidden Markov model or an autoregressive model. Bayes factors for samples and exhibits which had converging Metropolis-Hastings samplers were calculated using Chib and Jeliazkov's method (Chib and Jeliazkov (2001)). Bayes factors for samples and exhibits with non-converging Metropolis-Hastings samplers were calculated using Monte Carlo integration. Further details are given in Section 4.4 (and Appendix B). In total, 40 crime exhibits of a total of 70, and 94 samples of a total of 193 from general circulation were fitted with the hidden Markov model. The autoregressive model was used for the remainder.

To obtain the likelihood ratio for a seized sample, (4.7) was evaluated for each of the two propositions, H_C and H_B, with the individual integrals estimated using (4.3) and (4.6). This was done using the process described in Section 6.3.1. For the hidden Markov model, 29 of 263 seized samples required a repeat calculation of the likelihood ratio, using more posterior draws. The ranges of the five likelihood ratios obtained for those seized samples which required a repeat calculation are presented in table 6.3. As can be seen from the table, some of these ranges are large, an example is exhibit 67. The likelihood ratios used to present the results are the average of the likelihood ratios obtained (so based on either 15,000 or 100,000 evaluations for each integral).

As discussed in Section 6.3.1, methods for estimating confidence intervals for the estimated likelihoods were not practical due to long computation times. In practice, the range of likelihood ratio values calculated using the method described in Section 6.3.1 could serve as a measure of whether the estimate of the likelihood ratio is accurate. If the range of values calculated is too large, then the likelihood ratio should not be used to come to a conclusion about which of the two propositions (H_C or H_B) is supported by the evidence; more posterior draws are necessary to get a more accurate estimate. In the UK, likelihood ratios are often presented in court using a verbal scale, with a range of likelihood ratio values corresponding to a verbal measure of support. One such verbal scale can be seen in Evett, Jackson, Lambert and McCrossan (2000). The decision of the sorts of ranges of estimates of likelihood ratios that might be considered as too large could be made by reference to the verbal scale in use. If different estimates of the likelihood ratio are in different categories (and not just at the upper and lower edges of adjacent categories) then a subjective view should be taken as to whether a meaningful statement on the level of support given by the evidence to either proposition can be given. A likelihood ratio with a large range of estimates may be considered too unreliable to use as evidence.

Sample/ exhibit	Min LR obtained	Max LR obtained
13 (exhibit)	6.05	1.39×10^{1}
19 (exhibit)	9.61×10^{-2}	8.67
20 (exhibit)	2.38×10^{-1}	8.40×10^{-1}
38 (exhibit)	1.09×10^{2}	8.58×10^{2}
39 (exhibit)	3.21×10^{2}	2.44×10^{4}
40 (exhibit)	8.66×10^{1}	3.47×10^{2}
67 (exhibit)	1.19×10^{5}	1.44×10^{13}
68 (exhibit)	3.35×10^{1}	5.02×10^{4}
69 (exhibit)	8.41×10^{6}	8.23×10^{7}
17 (sample)	1.61×10^{-3}	1.95×10^{-3}
68 (sample)	9.71×10^{-5}	1.92×10^{-4}
72 (sample)	8.78×10^{-5}	1.95×10^{-3}
73 (sample)	4.23×10^{-5}	8.54×10^{-4}
78 (sample)	1.81×10^{-5}	1.52×10^{-4}
80 (sample)	4.47×10^{-5}	1.14×10^{-4}
87 (sample)	3.05×10^{-6}	7.17×10^{-6}
89 (sample)	7.69×10^{-6}	4.19×10^{-5}
90 (sample)	6.52×10^{-6}	2.15×10^{-5}
91 (sample)	1.33×10^{-6}	3.52×10^{-6}
92 (sample)	1.41×10^{-7}	3.96×10^{-7}
101 (sample)	1.12×10^{-2}	1.85×10^{-2}
119 (sample)	1.47×10^{-3}	3.10×10^{-3}
133 (sample)	1.19×10^{-3}	1.85×10^{-3}
158 (sample)	2.19×10^{-3}	9.90×10^{-3}
162 (sample)	4.21×10^{-5}	5.23×10^{-4}
167 (sample)	1.47×10^{-5}	5.48×10^{-5}
168 (sample)	4.80×10^{-6}	8.46×10^{-6}
172 (sample)	4.85×10^{-1}	2.27
193 (sample)	1.26×10^{-4}	2.65×10^{-4}

Table 6.3: Range of likelihood ratio values obtained for samples or exhibits requiring a repeat calculation of the likelihood ratio when treated as the seized sample (hidden Markov model).

6.3.4 Nonparametric model

Likelihood ratios for the nonparametric model (using both bandwidths) were evaluated using the method described in Section 4.5. The functions $\hat{f}_{D_i}$, used in the calculation of the likelihood ratio, were estimated as described in Section 6.2.4.

6.3.5 Standard model

Likelihood ratios for the standard model were evaluated using the method described in Section 4.6. As with the other models, each sample and exhibit was treated as the seized sample in turn. The remaining samples and exhibits (also excluding exhibits in the same case as the seized sample, if the seized sample was in set C) were used as training data sets B and C. Equation (4.10) was evaluated for each seized sample to obtain the likelihood ratio for that sample.

6.4 Results - the between sample distributions

6.4.1 Autoregressive model and hidden Markov model

Posterior distributions of all of the parameters for each of the 70 crime exhibits and 193 general circulation samples, were estimated for both the autoregressive model and the hidden Markov model. For the 193 general circulation samples, posterior distributions of the parameters $\theta^B_{A_i} = (\mu^B_i, (\sigma^B_i)^2, \alpha^B_i, \beta^B_i)$ and $\theta^B_{H_i} = (\mu^B_{1_i}, \mu^B_{2_i}, (\sigma^B_{1_i})^2, (\sigma^B_{2_i})^2, \alpha^B_{1_i}, p^B_{01_i}, p^B_{10_i}, \beta^B_{1_i}, \beta^B_{2_i})$, conditional on each sample $\mathbf{x}_i$ were estimated. For the 70 crime exhibits, posterior distributions of the parameters $\theta^C_{A_i} = (\mu^C_i, (\sigma^C_i)^2, \alpha^C_i, \beta^C_i)$ and $\theta^C_{H_i} = (\mu^C_{1_i}, \mu^C_{2_i}, (\sigma^C_{1_i})^2, (\sigma^C_{2_i})^2, \alpha^C_{1_i}, p^C_{01_i}, p^C_{10_i}, \beta^C_{1_i}, \beta^C_{2_i})$, conditional on each exhibit $\mathbf{y}_i$ were estimated. The parameters $\theta^B_{A_i}$ and $\theta^C_{A_i}$ correspond to the autoregressive model and the parameters $\theta^B_{H_i}$ and $\theta^C_{H_i}$ correspond to the hidden Markov model.

As described in Chapter 4, the posterior distributions for the individual samples and exhibits were combined to form overall between sample distributions for the parameters θ^B_A, θ^C_A, θ^B_H and θ^C_H. The between sample density functions for these parameters are given by $f(\theta^B_A \mid \mathbf{x})$, $f(\theta^C_A \mid \mathbf{y})$, $f(\theta^B_H \mid \mathbf{x})$ and $f(\theta^C_H \mid \mathbf{y})$ respectively. These overall between sample distributions can be thought of as mixture distributions, with definitions as given for the autoregressive model in (4.2). Figures 6.5, 6.6 and 6.7 give graphical representations of the marginal between sample density functions obtained for each parameter, based on the training data sets B and C. Figures 6.5 and 6.6 show the overall marginal between sample density function for each of the parameters in θ^C_H (left hand side) and θ^B_H (right hand side). Figure 6.7 shows the overall marginal between sample density functions for each of the parameters in θ^C_A (left hand side) and θ^B_A (right hand side). The between sample density functions relating to the hyperparameter β are not displayed, because estimates of this parameter were not required to evaluate the likelihood ratio. The autoregressive model parameters are used in both the autoregressive model, and the model which combines the autoregressive and hidden Markov models. A subset of the samples in B and exhibits in C are used to form the between sample density functions for θ^B_A and θ^C_A when the model combining the autoregressive and hidden Markov models is used. These subsets are formed from those samples and exhibits which have a Bayes factor that favours the autoregressive model over the hidden Markov model. As such, there are two different between sample density functions displayed for the parameters in θ^C_A and θ^B_A. The particular between sample distribution used to calculate the results depends on the model being used for the seized sample (either autoregressive or a mix of autoregressive and hidden Markov). In figure 6.7, the marginal between sample density functions used for the mixture of the autoregressive and hidden Markov models are shown with a dashed line and the marginal between sample density functions used for the autoregressive model alone are shown with a solid line. The between sample density functions shown with a solid line are estimated using all samples or exhibits in either B or C (those that converged). The between sample density functions shown with a dashed line are estimated using only those samples and exhibits which have a Bayes factor which favours the autoregressive model.

The graphs shown in figures 6.5, 6.6 and 6.7 are approximations of the between sample density functions actually used to evaluate the likelihood ratios. As mentioned in Section 3.3.6, the true marginal posterior density functions for the parameters μ_1, μ_2, σ_1^2, σ_2^2, p_{01} and p_{10}, conditional on each individual sample and exhibit, are bimodal. A rough rule of thumb was used to separate these bimodal density functions into unimodal density functions, so that, for example, μ_1 and μ_2 can be displayed separately. This was done by setting the smaller mean in each draw from the Metropolis-Hastings sampler as μ_1. Further approximations occur because in the diagrams shown, the posterior density functions relating to all samples and exhibits with converging Metropolis-Hastings samplers are included. When testing the models, the between sample distributions used to calculate the likelihood ratio did not include the posterior distribution relating to the sample or exhibit being treated as the seized sample (along with the posterior distributions relating to any other exhibits from the same case). If the models were used to evaluate likelihood ratios for an exhibit in a real case then the posterior distributions of all samples and exhibits (with converging samplers) would be included.

Means

As can be seen in figures 6.5, 6.6 and 6.7, the overall marginal between sample density function of the autoregressive mean of the general circulation samples (μ^B) occupies a similar range of values to the between sample density functions of the two lower means, μ_1^B and μ_1^C, of the hidden Markov model (for general circulation samples and crime exhibits respectively) and also the higher mean of the hidden Markov model for the general circulation samples (μ_2^B). All of these between sample density functions lie roughly between 5 and 7. Although there is a large overlap between the ranges of these four means, and the ranges of μ^C and μ_2^C (the autoregressive mean for crime exhibits and the higher mean of the hidden Markov model for crime exhibits), in general the between sample density functions of μ^C and μ_2^C are to the higher upper end of the between sample density functions of μ^B,μ_1^B, μ_1^C and μ_2^B (the ranges of the between sample density functions of μ^C and μ_2^C are roughly from 6 to 8). This is as expected: crime exhibit banknotes should have higher quantities of contamination than general circulation banknotes, and the higher means within the hidden Markov models should be larger than the lower means. The twelve exhibits (out of the 70 crime exhibits) that an expert declared as contaminated (see Section 5.3.1) were found to have higher means than most of the other exhibits.

It is interesting to note that the between sample density function of the lower mean of the hidden Markov model for crime exhibits (μ_1^C) has a large mode at around 7. This is higher than the main modes for the means associated with general circulation samples (μ_B, μ_1^B and even μ_2^B). This implies that the exhibits might not be a mixture of bundles of banknotes which have contamination consistent with general circulation and bundles of banknotes which are highly contaminated. Instead, some exhibits are a mixture of bundles of banknotes which still have two different contamination levels, but both of these contamination levels are higher than the level of contamination that is generally seen in general circulation.

Figures 6.5, 6.6 and 6.7 show that the marginal between sample density functions of the means (particularly for crime exhibits) are very multimodal. This stems from the large variation in quantities of contamination found on different exhibits and samples of banknotes (e.g. see figures 5.12, 5.13 and 5.14). Increasing the size of the training data sets C and B would reduce this multimodality.

Variances

The marginal between sample density functions of the variances of the autoregressive process ($(\sigma^B)^2$ and $(\sigma^C)^2$) are similar to those of the hidden Markov model ($(\sigma_1^B)^2$, $(\sigma_2^B)^2$, $(\sigma_1^C)^2$ and $(\sigma_2^C)^2$). The between sample density functions of $(\sigma_1^B)^2$ and $(\sigma_2^B)^2$ have a slightly larger variance than those of $(\sigma_1^C)^2$ and $(\sigma_2^C)^2$, with the same observed for $(\sigma^B)^2$ in comparison to $(\sigma^C)^2$. This is slightly surprising; crime exhibits, with their associated mix of highly and minimally contaminated banknotes, might be thought to have larger error variances. However, as can be seen from figures 5.12, 5.13 and 5.14, in many cases the crime exhibits do seem to have smaller ranges than general circulation samples, albeit at higher quantities of contamination.

Autocorrelation parameters

The marginal between sample density functions of the autocorrelation parameters generally lie between 0 and 0.7. The between sample density functions associated with crime exhibits have a mode slightly above 0.5, and the between sample density functions associated with general circulation samples have a mode slightly below 0.5. In figure 6.7, the marginal between sample density function of the autocorrelation parameter shifts to the left when those samples and exhibits which are best fitted with a hidden Markov model are removed from the estimate (i.e. for the AR(1)/HMM mix model). This could imply that samples and exhibits which are made up of different bundles of banknotes, with different levels of contamination, also have higher autocorrelation.

Transition probabilities

The marginal between sample density functions of the transition probabilities seen in figure 6.6 are generally between 0 and 0.5 and have a mode at around 0.05. The probabilities of transferring from a low contamination level to a high contamination level (p_{01}^B and p_{01}^C) are generally higher than the probabilities of transferring from a high contamination level to a low contamination level (p_{10}^C and p_{10}^B) for both general circulation samples and crime exhibits.

The prior distribution for the transition probabilities was a truncated beta distribution, with parameters of 0.6 and 4. The density function of a beta distribution with parameters of 0.6 and 4 has a mode at around 0.02, and assigns most weight to values up to around 0.4. The position of the mode of the truncated beta density function used as the prior for the i-th sample or exhibit depends on the number of banknotes in that sample or exhibit, but the truncated beta prior density function for a sample with 100 banknotes has a mode at around 0.05. The hyperparameters of this beta prior

distribution were chosen as it was thought that banknotes of either high or low contamination would cluster together (i.e. so that the transition probabilities were less than 0.5). The marginal between sample density functions for the transition probabilities seen in figure 6.6 have a similar mode to the prior density function, but assign most of their weight to smaller transition probabilities, up to around 0.25, suggesting that banknotes with similar quantities of contamination are clustered together more than was implied with the prior distribution.

6.4.2 Autoregressive model with random effects

The posterior density functions $f(\theta^B_{A_r} \mid \mathbf{x})$ (general circulation samples) and $f(\theta^C_{A_r} \mid \mathbf{y})$ (crime exhibits) were estimated for the parameters $\theta^B_{A_r} = (\mu^B_\mu, \sigma^B_\mu, \gamma^B_V, \beta^B_V, \mu^B_\alpha, \sigma^B_\alpha)$ and $\theta^C_{A_r} = (\mu^C_\mu, \sigma^C_\mu, \gamma^C_V, \beta^C_V, \mu^C_\alpha, \sigma^C_\alpha)$. These parameters are the parameters of the between sample distributions defined in (3.7), (3.8) and (3.9). The marginal posterior density functions for each of these parameters can be seen in figures 6.3 and 6.4.

Means

The posterior mode of μ^C_μ is approximately 6.75, compared to a posterior mode of 6.4 for μ^B_μ. These are similar to the values seen for the standard autoregressive model means (μ^B and μ^C) in figure 6.7. The posterior modes for σ^C_μ and σ^B_μ are approximately 0.5 (crime) and 0.37 (general circulation). The between sample plots for μ_C and μ_B in figure 6.7 have ranges of roughly $(6,8)$ and $(5,7)$ respectively, which would imply an approximate standard deviation of around 0.5; this is consistent with the standard deviation of 0.5 for crime exhibits found using the autoregressive model with random effects, and is slightly larger than the standard deviation of 0.37 found for general circulation samples.

Variances

The posterior modes for the parameters γ^B_V and β^B_V are around 14 and 2. These values would give a density function for $(\sigma^B)^2$ (inverse gamma) with a mode roughly equal to 0.15 and a range from around 0.1 to 0.3. This is similar to the between sample density function seen in figure 6.7 for $(\sigma^B)^2$. Similarly, the posterior modes for γ^C_V and β^C_V are around 6 and 0.4, corresponding to a density function for $(\sigma^C)^2$ with a mode just under 0.1 and a range between 0 and 0.2. This is also similar to the between sample density function of $(\sigma^C)^2$ seen in figure 6.7. A slightly larger variance for general circulation samples in comparison to crime exhibits was seen in Section 6.4.1 for the autoregressive model without random effects and the hidden Markov model. This larger variance is also seen with the autoregressive model with random effects.

Autocorrelation parameters

The posterior mode of μ^C_α is around 0.45 and the posterior mode of μ^B_α is slightly smaller, at around 0.42. This value for μ^B_α is similar to that found for α^B in the standard autoregressive model. However,

the mean of the between sample distribution for α^C (which is around 0.6) seems to be slightly higher than the value for μ_α^C of 0.45.

Summary

Overall, the marginal between sample density functions for the autoregressive model with random effects occupy similar ranges and have similar modes to those of the standard autoregressive model. There are two main differences. The first is that the variance of the between sample distribution of the mean for general circulation samples is slightly smaller for the model with random effects. The second is that the mode of the between sample density function of the autocorrelation parameter for the crime exhibits is larger for the standard autoregressive model.

6.5 Results - assessing the models

6.5.1 Rates of misleading evidence

Likelihood ratios were calculated for each model, treating each sample in B and each exhibit in C as the seized sample in turn, using the methods described in Section 6.3. Rates of misleading evidence were calculated for each of these models; these are displayed in table 6.4. A crime exhibit is said to provide misleading evidence if, when treated as the seized sample, the likelihood ratio associated with it is less than one, and so provides support for H_B. A general circulation sample is said to provide misleading evidence if, when treated as the seized sample, the likelihood ratio associated with it is greater than one, and so provides support for H_C. The rates of misleading evidence for crime exhibits and general circulation samples are given by the proportion of times that exhibits and samples respectively provide misleading evidence.

The two nonparametric models have the smallest rates of misleading evidence for crime exhibits. The method with adaptive bandwidth has a slightly smaller rate of misleading evidence than the method with a fixed bandwidth. These rates are, however, quite large; 25.7% of the evidence provided by crime exhibits is said to be misleading for the nonparametric model with adaptive bandwidth. The standard model has the largest rate of misleading evidence for crime exhibits, at 50%. As discussed in Section 5.3.3, some of the crime exhibits are not contaminated any more than general circulation samples (and are therefore in set C'), and yet are still included in the test (also see figure 5.15). As a result, low rates of misleading evidence are not expected for crime exhibits, even for well fitting models, and hence these rates should not be used for model comparison. As shown in Section 5.3.3, the rate of misleading evidence could depend more on the proportion of crime exhibits that are in the set C' than on the performance of the model.

The rates of misleading evidence for general circulation samples are smaller; the best is achieved with the hidden Markov model, which had a rate of misleading evidence of 10.4%. The autoregressive model also performed well, with a rate of 15.5%, similar to the rate of 16.1% for the autoregressive

	Hidden Markov Model	AR(1)	AR(1) random effects	Non-parametric fixed bw	Non-parametric adaptive nn	Standard model
Crime exhibit	0.357 (25/70)	0.371 (26/70)	0.386 (27/70)	0.271 (19/70)	0.257 (18/70)	0.500 (35/70)
General circulation	0.104 (20/193)	0.155 (30/193)	0.161 (31/193)	0.321 (62/193)	0.269 (52/193)	0.135 (26/193)

Table 6.4: Rates of misleading evidence, estimated as (r/n), where r is the number of samples or exhibits out of n analysed for which the likelihood ratio is said to be misleading in each context.

model with random effects. The standard model had a rate of misleading evidence of just 13.5% of samples, which is higher than the hidden Markov model, and slightly lower than the two autoregressive models. The nonparametric models performed badly here, with 32.1% and 26.9% of samples said to be misleading for the fixed bandwidth and adaptive bandwidth models respectively.

Using only the rates of misleading evidence for general circulation samples to assess the models does not take into account how well the models are performing with crime exhibits. As an example of the problems that this can cause, consider a model that gives likelihood ratios of less than one for all seized samples. This model would have a rate of misleading evidence of 0% for general circulation samples, but would be useless for the purposes of evidence evaluation. As mentioned in Section 5.3.1, all of the crime exhibits were analysed by forensic experts prior to being used as evidence in a criminal case. Twelve of the exhibits were declared as contaminated by these experts. The log-likelihood ratios obtained for these twelve exhibits are given in table 6.5. As can be seen, all of the models returned log-likelihood ratio values of greater than zero (corresponding to likelihood ratio values of greater than one) for all twelve of these exhibits, suggesting agreement with forensic experts.

The rates of misleading evidence in table 6.4 show which of the methods is best at minimising the number of occasions on which a piece of evidence is found to be misleading, but they do not give an indication of the size of the likelihood ratios obtained for each of the models. In practice, a likelihood ratio value close to one does not provide strong evidence. A method which gives larger likelihood ratios, where these large values are justified, is of more value when being used in evidence evaluation. The plots in figures 6.8, 6.9 and 6.10 consider the actual values of the likelihoods and the likelihood ratios for each model; these plots are discussed in the following two sections.

6.5.2 Tippett plots

The Tippett plots in figure 6.8 indicate the proportion of general circulation samples (dashed) and crime exhibits (solid) that have a log-likelihood ratio greater than the horizontal axis of the plot. Figure 6.8 shows that the autoregressive model of order one has log-likelihood ratios which are much closer to the neutral value of zero, for both crime exhibits and general circulation samples, than the hidden Markov model and nonparametric models, meaning that it is less useful for discriminating between crime exhibits and general circulation samples. The autoregressive model with random effects and the standard model have larger log-likelihood ratios than the autoregressive model without random

Exhibit number	HMM	AR(1)	AR(1) - random effects	nonparametric fixed bandwidth	nonparametric variable bandwidth	Standard model
1	7.37	6.05	5.45	31.02	39.02	32.61
3	3.51	3.67	3.74	5.19	6.43	4.68
16	6.61	7.51	11.31	6.92	7.14	2.89
23	7.51	6.32	5.73	8.64	7.64	7.72
38	5.38	6.64	8.87	11.61	12.55	7.39
39	7.31	10.39	16.12	20.43	22.69	8.51
40	4.91	2.24	7.62	0.05	21.53	0.60
42	4.35	4.09	6.27	6.23	8.03	2.47
43	6.89	7.06	10.73	6.80	8.61	2.06
57	4.66	3.58	2.34	6.24	11.13	5.45
67	16.52	0.57	6.05	244.80	262.25	7.51
69	17.42	0.48	5.96	128.69	169.64	5.44

Table 6.5: Log likelihood ratios of exhibits declared as contaminated by an expert

effects for crime exhibits that are not misleading, but have log-likelihood ratios similar in size (so close to zero) to those seen for the autoregressive model without random effects for general circulation samples.

The two nonparametric models give larger absolute log-likelihood ratio values than all of the parametric models, but in some cases these large values are misleading because some general circulation samples have very large and positive log-likelihood ratios. The standard model also gives a large and positive misleading log-likelihood ratio to one of the general circulation samples. The hidden Markov model and both of the autoregressive models have small log-likelihood ratios for general circulation samples that provide misleading support, which means that seized samples that are from general circulation are unlikely to provide strongly misleading evidence with these models.

The two exhibits which, when treated as the seized sample, have the largest log-likelihood ratios using the hidden Markov model are exhibits 67 and 69 (see table 6.5). Exhibit 69 had a slightly unstable log likelihood ratio estimate, with estimates ranging from 16 to 18 (using the results in table 6.3), and exhibit 67 had a very unstable log-likelihood ratio estimate, with estimates ranging from 12 to 30. Therefore, the large tail on the Tippett plot for crime exhibits when the hidden Markov model is used to calculate likelihood ratios may be misleading. However, log-likelihood ratio values for other exhibits do still seem to be slightly larger when the hidden Markov model is used to calculate likelihood ratios, in comparison to the autoregressive model without random effects, and log-likelihood ratio values for general circulation samples are much smaller, which suggests that the model is discriminating between the two sets of data more effectively than the autoregressive model without random effects.

6.5.3 Scatter plots

In the scatter plots in figures 6.9 and 6.10, the samples and exhibits are plotted according to their log likelihood values under the prosecution and defence propositions, H_C and H_B. Crime exhibits that, when treated as the seized sample, have not provided misleading evidence are above the solid line and

general circulation samples that, when treated as the seized sample, have not provided misleading evidence are below the solid line. In figure 6.10, outliers with very small log-likelihoods under one or both propositions have been removed.

In figure 6.10, it can be seen that log-likelihood values cluster around the line in the case of the autoregressive model (both with and without random effects), suggesting equal values under both propositions, are slightly further from the line for the hidden Markov model and are furthest from the line for the nonparametric models. This corresponds to conclusions drawn from the Tippett plots, that the hidden Markov model and the nonparametric models provide the largest absolute values of log-likelihood ratios.

Nonparametric model outliers

The scatter plots for the nonparametric models in figure 6.9 have two concerning features. The first concerns the four crime exhibits, indicated with a ∘ in the bottom left corner of the nonparametric model plots in figure 6.9, that have much larger log-likelihood values under proposition H_C than under proposition H_B. For the parametric models (excluding the standard model) these four exhibits have log-likelihood values under proposition H_C which are similar in value to their log-likelihood values under proposition H_B. There is a large difference in the log-likelihood values generated for these four exhibits by the nonparametric models and the parametric models. These four exhibits contain 1023, 1030, 1099 and 1065 banknotes, and they all come from the same case. The next nearest in size of all exhibits contains 606 banknotes. Thus, these outliers could be indicative of a lack of robustness in the nonparametric methods for samples or exhibits with a large number of banknotes. The conditional density function estimates, evaluated at each pair of banknotes, are multiplied together to obtain an overall likelihood. Since these four exhibits have large numbers of banknotes, if there were small errors in the estimate of the conditional density function for each pair of banknotes, then taking the product of these small errors would result in a large overall error. An alternative explanation for the lack of robustness of the nonparametric methods for these four large exhibits might be that these exhibits contain some banknotes with contamination quantities that are very unusual (this is likely, given their size). Then, the estimated conditional density functions $\hat{f}_{D_i}$ will need to be evaluated in their tails for these unusually contaminated banknotes. As discussed in Section 3.4, there are problems associated with the use of kernel density estimates for the estimation of the tails of density functions. These problems could be the cause of the large errors seen for these four exhibits. It is interesting to note that the problems with these four exhibits are seen for both the fixed and variable bandwidth models, so the use of a variable bandwidth does not seem to have resolved the problem for these data.

Problems with a lack of robustness of the nonparametric methods for large exhibits can also be seen in table 6.5. Exhibits 67 and 69 have 1099 and 1023 banknotes respectively, and the log-likelihood ratios for these exhibits are very different to the values obtained using parametric models. This problem could be investigated further if more data could be obtained for the training data sets.

The second concerning feature of the scatter plots for the nonparametric models is that some general circulation samples have large log-likelihood values under proposition H_C in comparison to their log-likelihood values under proposition H_B when the nonparametric models are used. These can be seen in figure 6.10. Again, this could be due to problems with a lack of robustness of the nonparametric models when estimates are required in the tails of conditional density functions.

The standard model

The scatter plots for the standard model in figures 6.9 and 6.10 look very different to the scatter plots for the other models. The log-likelihood values are much smaller in absolute value than those of the other models, especially considering that different scales have been used on the axes. Also, with the exception of some crime exhibits and one general circulation sample, most samples and exhibits cluster in one small area rather than extending diagonally across the plot. The absolute sizes of the log-likelihoods are small because the means of the measurements for each sample and exhibit are used directly in the calculation of the likelihoods, rather than the individual measurements on each banknote, as done with the other models. As such, the number of banknotes in each sample or exhibit does not have as big an effect on the log-likelihood values, and so the points do not extend along a diagonal line, as seen with the other models.

As can be seen in the scatter plot for the standard model in figure 6.10, there are some outlying crime exhibits which extend in a roughly horizontal line across the plot. These exhibits all have roughly similar log-likelihoods under H_C to the other samples and exhibits, but have smaller log-likelihoods under H_B. Figure 5.15 is a graphical representation of the two between sample density functions used in the standard model (one for crime exhibits and one for general circulation samples). The between sample density function for the crime exhibits has two modes. One of these modes is in the same position as the mode for general circulation samples, at around 6.5, and one is larger, at around 7. The outlying crime exhibits in the scatter plot for the standard model all have means greater than 7.3, meaning that they have mean contamination roughly in line with the larger mode of the between sample density function of the crime exhibits (so have large likelihoods under H_C). However, these outlying exhibits have means which are very different to the means of general circulation samples (so have small likelihoods under H_B). The remaining samples and exhibits have means that are smaller than 7.3, which is consistent with samples from general circulation, but also consistent with a substantial proportion of crime exhibits, and so they all have likelihood ratios close to one. The effect of this bimodal between sample density function can also be seen in the Tippett plot for the standard model, in figure 6.8. Most of the general circulation samples and some of the crime exhibits have log-likelihood ratios very close to zero; these are the samples and exhibits with contamination consistent with the lower mode of the between sample density function of the crime exhibits in figure 5.15.

In the scatter plot for the standard model in figure 6.9, there is one general circulation sample

which is very strongly misleading. The log-likelihood ratio for this sample is 22.6, which is extremely large in comparison to other samples and exhibits, and hence is concerning as it is misleading. As a comparison, the log-likelihood ratio for the same sample using the autoregressive model is 3.2 and the log-likelihood ratio for the sample sample using the hidden Markov model is -5.2. This misleading general circulation sample is also seen in the long tail to the right of the Tippett plot for the standard model in figure 6.8.

6.5.4 A comparison of the standard model to models accounting for autocorrelation

The standard model performs well in terms of the rate of misleading evidence. Fewer general circulation samples when used as the test set are said to be misleading than for the autoregressive models, and only 3% more are said to be misleading than when the hidden Markov model is used. The standard model is also considerably easier to implement. There are, however, problems associated with this model, one of which, the very misleading general circulation sample, was discussed in the previous section.

Another problem with the standard model is that autocorrelation is not taken into account. It was shown in Section 5.6.2 that autocorrelation was present in the majority of samples and exhibits. One result of ignoring autocorrelation is that likelihood ratios may be overstated. The possible effects of this can be seen in table 6.5. Exhibits 1 (46 banknotes), 3 (32 banknotes), 23 (34 banknotes), 38 (21 banknotes) and 57 (21 banknotes) are the five smallest exhibits of the twelve said to be contaminated by an expert. All of these exhibits have larger log-likelihood ratios for the standard model, in comparison to the other three parametric models (with the exception of exhibit 38, for which the autoregressive model with random effects has a larger log-likelihood ratio). This is particularly true in the case of the first exhibit. The standard model uses only the mean of the seized sample to calculate its likelihood ratio. All of the exhibits mentioned above have high means. A small group of highly contaminated banknotes which are sequentially close together within one of these small exhibits could have a large influence on the mean. When autocorrelation is accounted for, the influence of this small group of highly contaminated banknotes is reduced. This is because the high quantities of contamination on this small group of banknotes could be due to just one or two highly contaminated banknotes having transferred cocaine onto the surrounding banknotes.

Another explanation for these large likelihood ratios for small exhibits when autocorrelation is not accounted for might be the choice of the weights used in the calculation of the likelihood ratio for the autoregressive model (without random effects) and the hidden Markov model. The choice of weights was found to have some effect on likelihood ratios for small seized samples. However, this choice did not fully account for the differences found between the results for the standard model and the results for the other parametric models. Further details are given in Section 6.5.7.

6.5.5 Results in relation to the modelling of different levels of contamination on different bundles of banknotes

The hidden Markov model has the smallest rate of misleading evidence for general circulation samples. This rate is, however, not a great deal smaller than the rates of the two autoregressive models and the standard model. The hidden Markov model gives larger likelihood ratio values than the autoregressive model without random effects for some seized samples known to be crime exhibits, which makes it more useful for discrimination in practice. However, these likelihood ratio values are of a similar size to those seen for the autoregressive model with random effects and the standard model.

One benefit of the use of the hidden Markov model is that it allows for different bundles of banknotes having different levels of contamination. A seized sample with several bundles of banknotes contaminated at a high level, and several contaminated at a low level would have a mean which was not out of the ordinary; as a result, models which allow for only one mean level would assign a low likelihood ratio to such a seized sample. The hidden Markov model, however, allows for two different levels of contamination to be taken into account. As an example, exhibit 69 in table 6.5 has 1023 banknotes, of which 55 have a log peak area of greater than 8, but 185 have a log peak area of less than 7. Exhibit 67 has 1099 banknotes, of which 79 have a log peak area of greater than 8, but 205 have a log peak area of less than 7. In total, there are 24,285 banknotes in the general circulation database and just 19 of these have a log peak area of greater than 8. Of the other exhibits declared as contaminated by experts, exhibit 1 has 20 banknotes with a log peak area of greater than 8 and exhibit 23 has one banknote with a log peak area greater than 8. The other exhibits in table 6.5 have no banknotes with log peak areas greater than 8. These numbers imply that exhibits 67 and 69 are highly unusual in comparison to the general circulation database, despite having 205 banknotes and 185 banknotes respectively which have a log peak area of less than 7 (which is consistent with general circulation). These two exhibits also have more highly contaminated banknotes than the other exhibits which were declared as contaminated by experts.

The hidden Markov model assigns the largest log-likelihood ratios to exhibits 67 and 69, values of 16.52 for exhibit 67 and 17.42 for exhibit 69. These log-likelihood ratios are substantially higher than those assigned by the hidden Markov model to the other exhibits in table 6.5. The standard model and the autoregressive model with random effects assign fairly large likelihood ratios to these exhibits, but they are not out of the ordinary in comparison to other exhibits declared as contaminated by experts. The autoregressive model without random effects assigns small likelihood ratios of just 0.57 for exhibit 67 and 0.48 for exhibit 69. The two autoregressive models and the standard model assign smaller likelihood ratios than warranted to these two exhibits because they cannot account for the mixture of banknotes with low contamination and banknotes with high contamination.

6.5.6 A comparison of the two autoregressive models, with and without random effects

The log-likelihood ratios of the crime exhibits seen in table 6.5 are similar for the hidden Markov model and the standard autoregressive model (with the exception of exhibits 67 and 69, discussed in the previous section). However, the log-likelihood ratios for the autoregressive model with random effects are larger than those for the standard autoregressive model for many of the exhibits; this can also be seen in figure 6.8, where the Tippett plot for the autoregressive model with random effects suggested that this model was assigning larger likelihood ratios to seized samples known to be from set C. The log-likelihood ratios for exhibits 16, 38, 39, 40, 42, 43, 67 and 69 particularly are a lot larger when the autoregressive model with random effects is used. Further investigation revealed that these exhibits have much smaller likelihoods of H_B when random effects are used. These smaller likelihoods of H_B could reflect the slightly smaller value of σ_μ^B seen for general circulation samples for the model with random effects, in comparison to the standard deviation of μ_B estimated from visual inspection of the between sample density function of μ_B for the standard autoregressive model (figure 6.7).

The two autoregressive models have similar rates of misleading evidence for general circulation samples, so the differences in log-likelihood ratios for the two models, for crime exhibits in table 6.5, lead to questions about which model is better in practice. One way to answer this is to consider the model fit. The numerators of the likelihood ratios for seized samples known to be in set C, evaluated for different models, can be used as a proxy for the model fit for crime exhibits. Similarly, the denominators of the likelihood ratios for seized samples known to be in set B, evaluated for different models, can be used as a proxy for the model fit for general circulation samples. If the seized sample $\mathbf{z}$ is known to be in set D, then the ratio

$$\frac{f(\mathbf{z} \mid H_D, M_z = M_A)}{f(\mathbf{z} \mid H_D, M_z = M_{A_r})}$$

gives the Bayes factor for the comparison of the models M_A (autoregressive) and M_{A_r} (autoregressive with random effects) for the data $\mathbf{z}$. This Bayes factor was calculated for all crime exhibits in set C and all general circulation samples in set B. Of the 70 crime exhibits, 51% had a Bayes factor which favoured the autoregressive model without random effects and of the 193 general circulation samples, 62% had a Bayes factor which favoured the autoregressive model without random effects. Therefore, there is a slight preference for the standard autoregressive model, particularly for general circulation samples. In Section 4.2, it was noted that the autoregressive model with random effects makes an assumption of Normality for the between sample distribution of the mean, an assumption which may not hold for the data considered here (see for example, the between sample plots for μ_C in figure 6.7).

6.5.7 The effect of the choice of weights

The weights v_i used for the autoregressive model (without random effects) and the hidden Markov model were taken to be equal to

$$v_i = \frac{n_{D_i}}{\sum_{i=1}^{m_D} n_{D_i}},$$

so that the weight for a sample or exhibit varies with the number of banknotes in that sample or exhibit. These weights give large samples or exhibits in the training data sets a greater influence on the between sample distribution, and hence the resulting likelihood ratio. In Section 6.5.4, it was shown that the standard model gave much larger log-likelihood ratios, when the five smallest exhibits that were declared as contaminated by experts were taken to be the seized sample, than both the autoregressive model and the hidden Markov model. These larger likelihood ratios could imply that the standard model overstates likelihood ratios for small seized samples, because the standard model does not account for autocorrelation. Alternatively it could be that the five small exhibits have similar patterns of contamination to one another, and that this pattern of contamination is different in some way to the larger exhibits. With weights that give more influence to larger exhibits, these differences in patterns of contamination would result in small likelihoods of H_C for small seized samples (because the small seized samples are unlike the large exhibits in set C). To test this, likelihood ratios were calculated with weights equal to

$$v_i = \frac{1}{m_D}$$

so that each sample or exhibit in each training data set had equal weight, regardless of the number of banknotes in the sample or exhibit. The rates of misleading evidence obtained for the hidden Markov model and the autoregressive model (without random effects) for these new weights can be seen in table 6.6. The rate of misleading evidence for general circulation samples for the hidden Markov model is the same regardless of the choice of weights, but the rate for general circulation samples for the autoregressive model reduces from 16% to 11% when the weights are changed to $1/m_D$. This drop of 5% corresponds to nine general circulation samples which, when treated as the seized sample, have a likelihood ratio which is misleading for uneven weights but have a likelihood ratio which is not misleading for weights of $1/m_D$. When treated as the seized sample, these nine general circulation samples have likelihood ratios which, although misleading, are very close to one for uneven weights. The largest of these nine likelihood ratios is 2.26. Therefore, changing the weights did not result in large changes in the likelihood ratios associated with these samples.

The log-likelihood ratios associated with weights of $1/m_D$ for the twelve exhibits declared as contaminated by experts can be seen in table 6.7. It can be seen that changing the weights has not changed the log-likelihood ratios by a large amount for either of the models, with the exception of for the first exhibit. Some of the five small exhibits discussed in Section 6.5.4 (1,3,23,38 and 57)

	Hidden Markov Model	AR(1)
Crime exhibit	0.329 (23/70)	0.300 (21/70)
General circulation	0.104 (20/193)	0.109 (21/193))

Table 6.6: Rates of misleading evidence when weights are equal to $1/m_D$, estimated as (r/n), where r is the number of samples or exhibits out of n analysed for which the likelihood ratio is said to be misleading in each context.

have slightly larger log-likelihood ratios when the weights are changed (in particular exhibit 3 for the autoregressive model, exhibit 23 for both models, and exhibit 38 for the hidden Markov model) but these log-likelihood ratios are still smaller than those of the standard model, for all but exhibit 23 for the hidden Markov model. This is evidence to suggest that, although the choice of weights seems to have a small effect on the log-likelihood ratios for small exhibits, this choice does not fully account for the larger log-likelihood ratios seen when the standard model is used to evaluate the likelihood ratio for these small exhibits. This increase in log-likelihood ratio could instead be caused by the assumption of independence made by the standard model.

Exhibit number	HMM	HMM - new weights	AR(1)	AR(1) - new weights	Standard model
1	7.37	4.59	6.05	3.35	32.61
3	3.51	3.26	3.67	4.24	4.68
16	6.61	6.65	7.51	6.98	2.89
23	7.51	8.39	6.32	7.18	7.72
38	5.38	6.35	6.64	6.92	7.39
39	7.31	8.38	10.39	9.86	8.51
40	4.91	5.26	2.24	1.55	0.60
42	4.35	4.67	4.09	4.38	2.47
43	6.89	7.11	7.06	6.90	2.06
57	4.66	4.89	3.58	3.12	5.45
67	16.52	17.51	0.57	1.39	7.51
69	17.42	17.19	0.48	1.13	5.44

Table 6.7: Log likelihood ratios of exhibits declared as contaminated by an expert. The new weights are $v_i = 1/m_D$.

6.5.8 Summary of results

Overall, based on the rates of misleading evidence for general circulation samples, and the problems with outliers, it is recommended that the parametric models are used to calculate likelihood ratios for data relating to traces of cocaine on banknotes. The hidden Markov model has a slightly better rate of misleading evidence for general circulation samples than the autoregressive model (both with and without random effects) and the standard model, and discriminates slightly better than the autoregressive model by returning larger absolute log-likelihood ratios when samples and exhibits are not said to be misleading. These larger absolute values of log-likelihood ratios make the hidden Markov

model more useful than the autoregressive model for discriminating between general circulation samples and crime exhibits in practice.

The hidden Markov model allows for two different levels of contamination within samples and exhibits, as well as accounting for autocorrelation between measurements on adjacent banknotes. By not accounting for autocorrelation, there is a risk that likelihood ratios will be overstated. By not allowing for different levels of contamination on different bundles within samples and exhibits, there is potential for understating likelihood ratios. However, care should be taken when using the hidden Markov model, as the Monte Carlo error can be large, so (particularly with large seized samples) the estimates of the likelihood ratio are subject to a greater variance than those of the autoregressive models and the standard model.

6.6 Dissemination of methods - Graphical User Interface

A graphical user interface (GUI) was developed so that two of the parametric models developed in this thesis (the hidden Markov model and the autoregressive model without random effects) can be used to evaluate evidence relating to traces of cocaine on banknotes in practice. Draws from the posterior distributions of the parameters associated with each sample and exhibit in each of the two training data sets, for both of the models, were stored as part of the GUI. Storing these draws means that the Metropolis-Hastings sampler does not need to be implemented by the user. The peak detection algorithm in Section 5.4 was programmed in R, using the MassSpecWavelet package (Du et al. (2006)), and the likelihood ratio calculations in Chapter 4 were programmed in `C++`, using libraries from the Rcpp R package (Eddelbuettel and Francois (2011); Eddelbuettel (2013)).

To use the GUI, the raw data relating to the seized sample ($\mathbf{z}$) should be loaded into the GUI using the menu bar. The user can then input details relating to the scan numbers of the standard injections and the way in which the individual runs combine to form exhibits. The peak detection algorithm can then be implemented; the user is able to modify the positions of the peaks so that adjustments can be made if peaks have been falsely identified as banknotes or have not been detected at all. A screenshot from the GUI after the peak detection algorithm has been run is given in figure 6.11. The raw data are shown on the graph with a solid line, the detected peaks are shown with a dark coloured circle and the detected standards are identified with a light coloured circle. Details relating to the exhibits associated with the raw data (some of which is entered by the user and some of which is obtained from the raw data file) are shown in the table above the graph (the case number has been removed from the screenshot for the purposes of anonymity).

Using the peak areas obtained from the implementation of the peak detection algorithm, the user can calculate likelihood ratios using both the hidden Markov model and the autoregressive model. The hidden Markov model used is that discussed in Section 4.4, which allows for some samples and exhibits to be modelled with a hidden Markov model, with the remainder modelled using an autoregressive process. For the reasons given in Section 6.5.8, it is recommended that the results from

the hidden Markov model are used to come to a decision about the seized sample. A screen shot of the GUI after likelihood ratio calculations have been carried out is given in figure 6.12. The trace plot of the logarithms of the peak areas of the seized sample is shown on the right; the log peak areas are listed on the left. The user can use the trace plot to identify any errors that might have occurred in the calculation of the peak areas. The table at the bottom of figure 6.12 gives the likelihood ratio results for the seized sample. As discussed in Section 6.3, the likelihood ratio calculations are repeated; the range of values obtained is outputted and displayed in the table. The user can check the ratio between the maximum and the minimum likelihood ratios obtained. If this ratio is too high, the likelihood ratio can be re-calculated using 20,000 draws from the posterior distributions, instead of 5,000. This more accurate calculation incurs an increase in calculation time.

6.7 Conclusion

In this chapter, the models developed in Chapter 3 were used to evaluate likelihood ratios for evidence relating to traces of cocaine on banknotes. The models were an autoregressive model of order one (both with and without random effects), a hidden Markov model and a nonparametric model, with two different methods of bandwidth selection. All of these models were designed to take autocorrelation between adjacent banknotes into account. The hidden Markov model also accounted for two different levels of contamination on different bundles within the same sample or exhibit. The data described in Chapter 5 were used to test and compare the models, and comparisons were made with a standard model which assumes independence between measurements on adjacent banknotes. This standard model is described in Section 4.6.

Rates of misleading evidence, Tippett plots and scatter plots were used to evaluate the performance of each of the models. The hidden Markov model was found to have the smallest rate of misleading evidence for general circulation samples, with a rate of 10.4 %. The autoregressive models and the standard model had similar rates of misleading evidence for general circulation samples, ranging between 13.5% and 16.1%. The nonparametric models performed poorly, with over 25% of samples said to provide misleading evidence. Analysis of the Tippett and scatter plots suggested that the nonparametric models were not robust for samples and exhibits with a large number of banknotes. In addition, these plots showed that the hidden Markov model was giving slightly larger absolute values of log-likelihood ratios than the autoregressive models, suggesting that the hidden Markov model was discriminating between general circulation samples and crime exhibits slightly more effectively. Consideration of the likelihood ratios obtained for twelve crime exhibits which were declared by forensic experts as contaminated suggested that the standard model might be overstating the likelihood ratios for small seized samples. This overstating of the likelihood ratios could be because autocorrelation was not accounted for. The likelihood ratios of two of the crime exhibits declared as contaminated by experts suggested that by not allowing for two different levels of contamination, likelihood ratios for exhibits which are a mixture of very highly contaminated banknotes and banknotes with low

contamination might be understated. Overall, it was recommended that the hidden Markov model is used for analysis of these data.

It was shown in Section 5.3.3 that it is not possible to obtain low rates of misleading evidence for crime exhibits, even with an accurate model, because of the large number of crime exhibits which have contamination consistent with samples from general circulation. The rates of misleading evidence obtained for crime exhibits for all of the models developed in this thesis support this view. If low rates of misleading evidence cannot be obtained for crime exhibits, then the approaches given here cannot be used to provide support for proposition H_B. This is because the models are known to provide misleading evidence in support of H_B a large proportion of times. These large rates of misleading evidence are not a reflection on the accuracy of the models and hence cannot be used to assess the models.

Results in this chapter demonstrate that the parametric models developed can be used to evaluate likelihood ratios for autocorrelated data, and also data which are driven by an underlying latent Markov chain. The results also show that it is possible to evaluate likelihood ratios for evidence relating to traces of cocaine found on banknotes, with respect to the two propositions H_B and H_C. For a method to be useful in practice for these data, it needs to have a low rate of misleading evidence for general circulation samples, so that samples from B are not falsely said to support proposition H_C. To be useful, the method must also be able to provide support for H_C for some seized samples known to be crime exhibits, with likelihood ratios for these seized samples large enough to give some level of support to H_C (i.e. so that likelihood ratios for crime exhibits are not all close to one). If the method did not do this, then there would be nothing gained from using cocaine traces on banknotes as evidence. The results presented here give three new methods, the autoregressive model, the autoregressive model with random effects and the hidden Markov model, which have small rates of misleading evidence for general circulation samples and which also give large likelihood ratios to some seized samples known to be crime exhibits in set C, so that both of these requirements are satisfied.

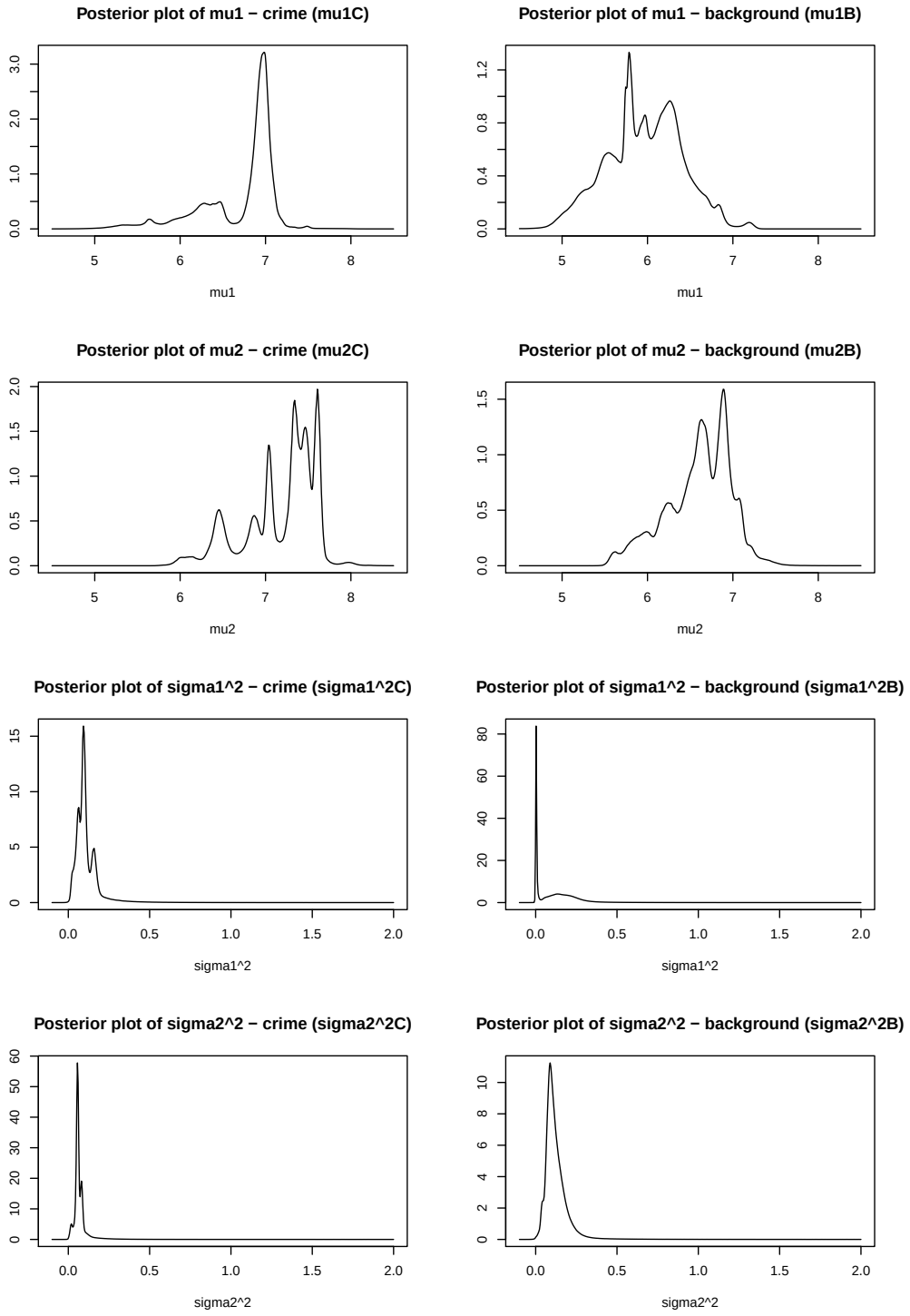

Figure 6.5: Plots of the marginal between sample density functions obtained for hidden Markov model parameters

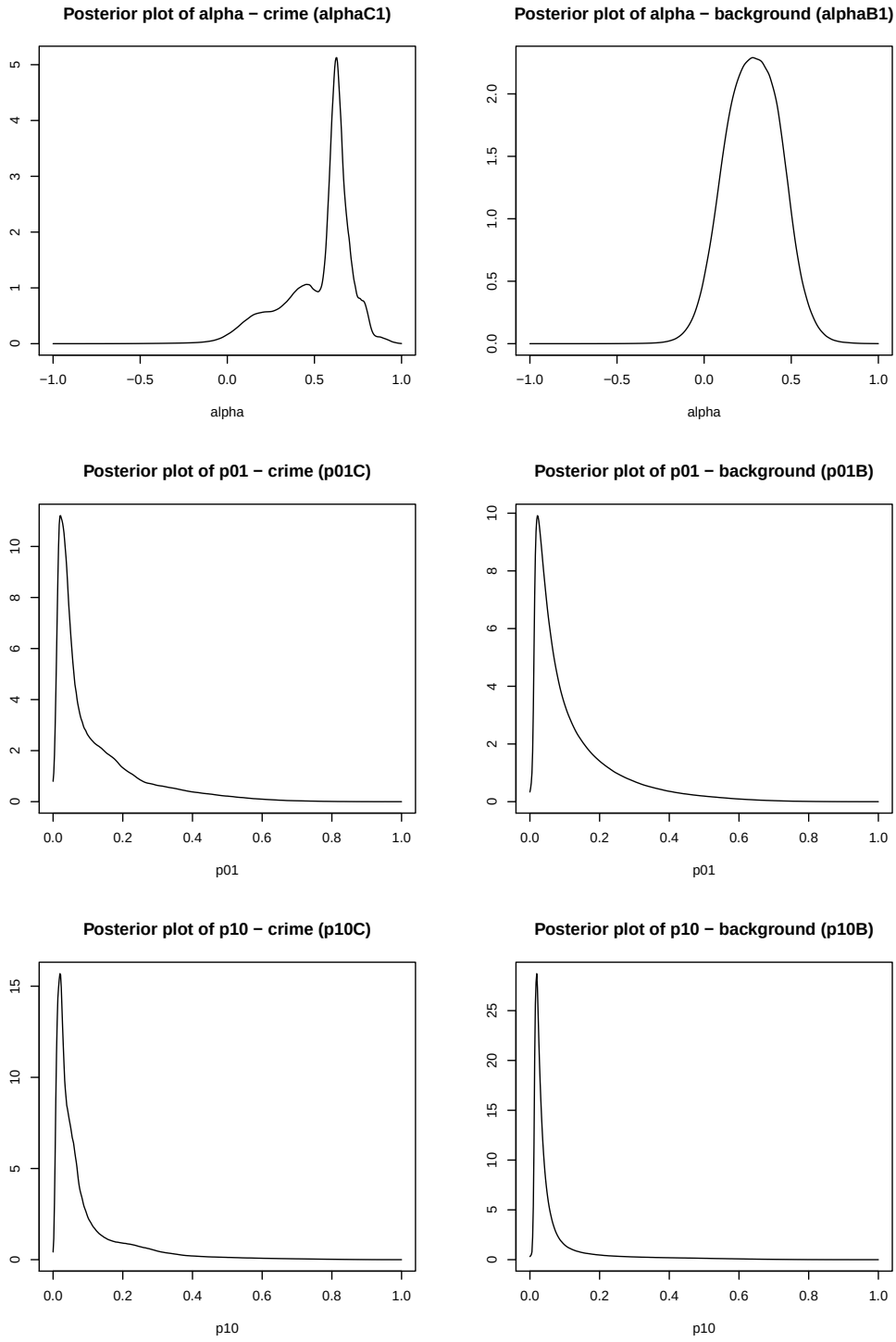

Figure 6.6: Plots of the marginal between sample density functions obtained for hidden Markov model parameters

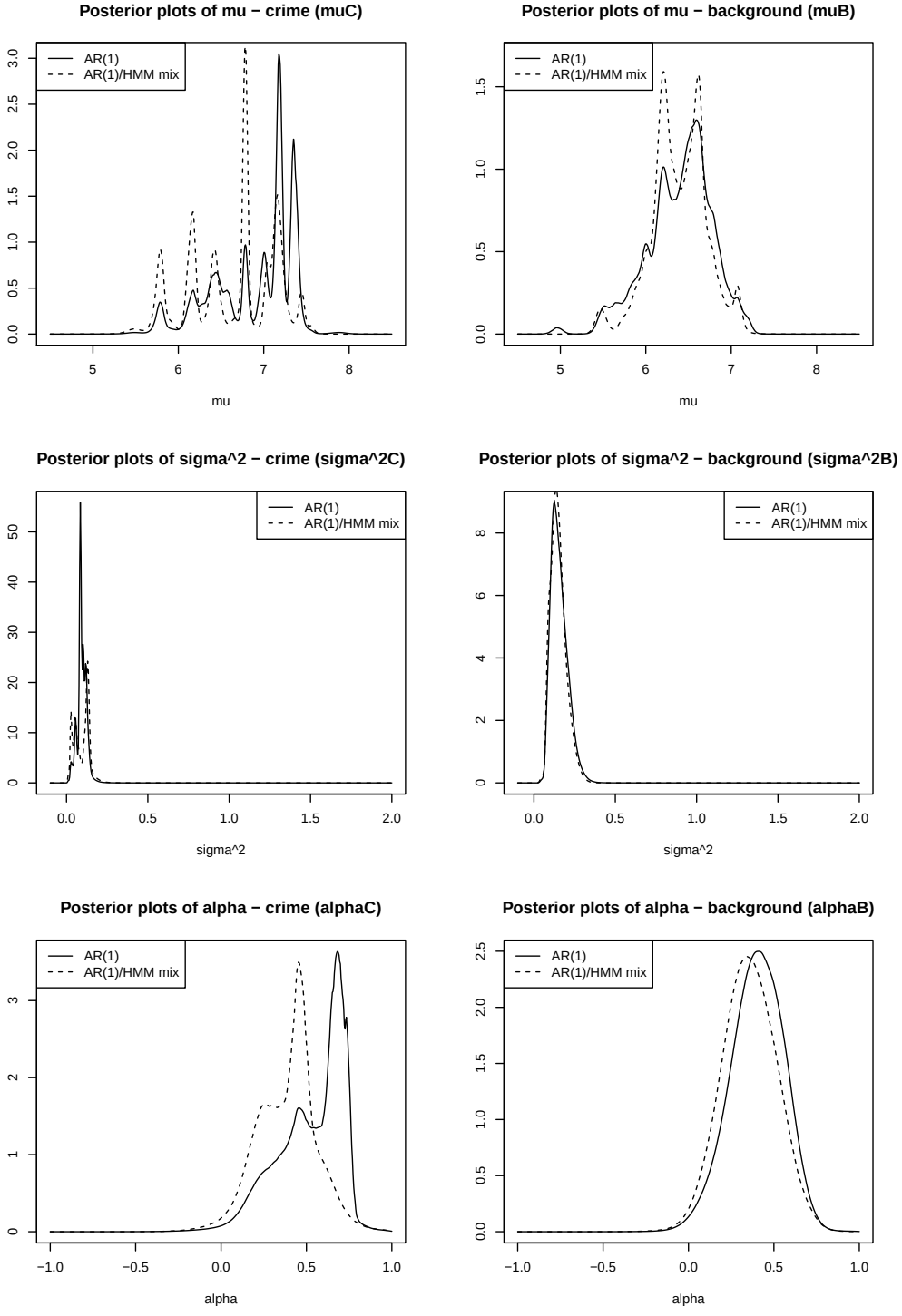

Figure 6.7: Plots of the marginal between sample density functions obtained for autoregressive model parameters

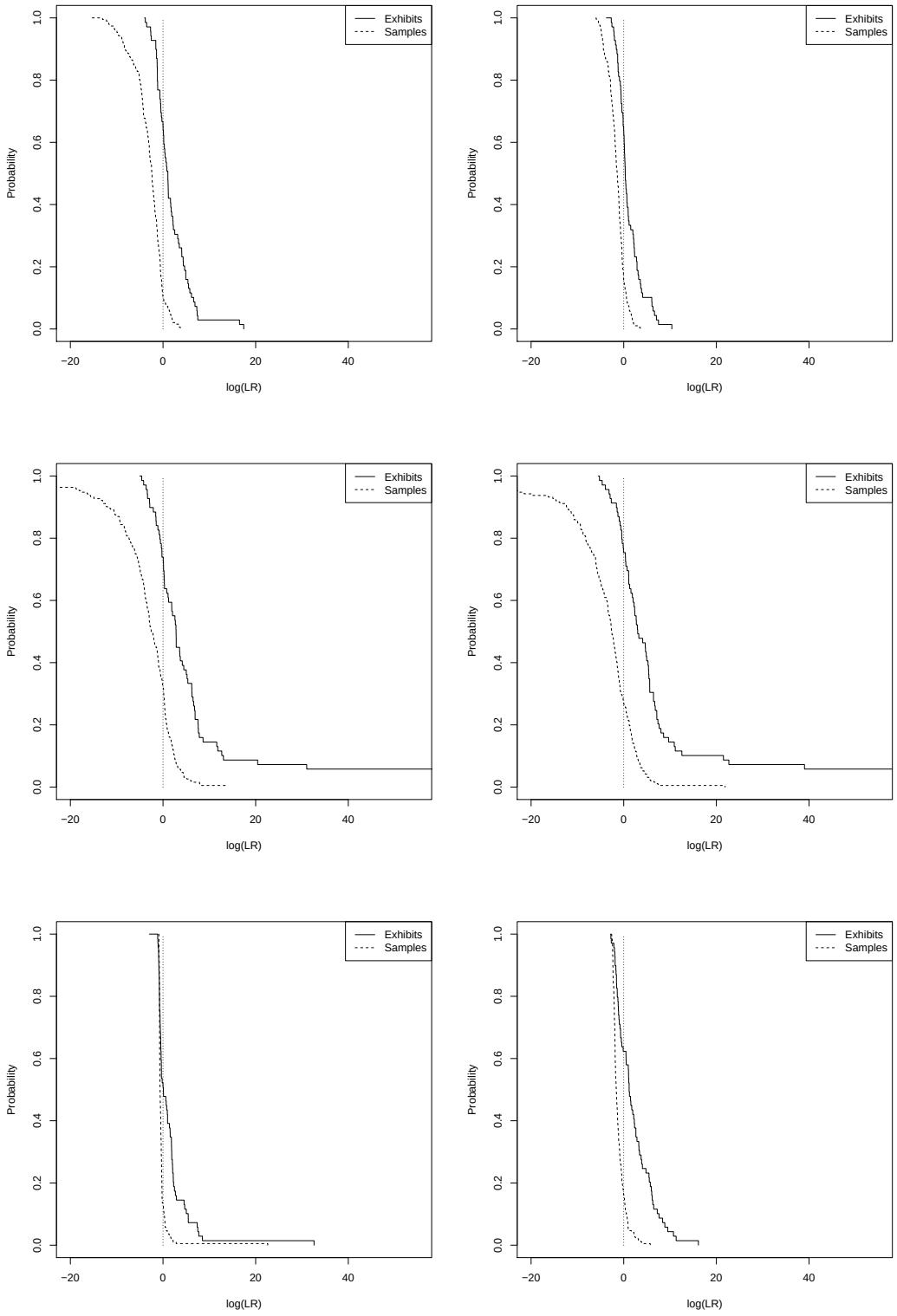

Figure 6.8: Tippett plots of likelihood ratio values. Clockwise from top left - hidden Markov model, AR(1), adaptive bandwidth, AR(1) with random effects, standard model (assuming independence), fixed bandwidth

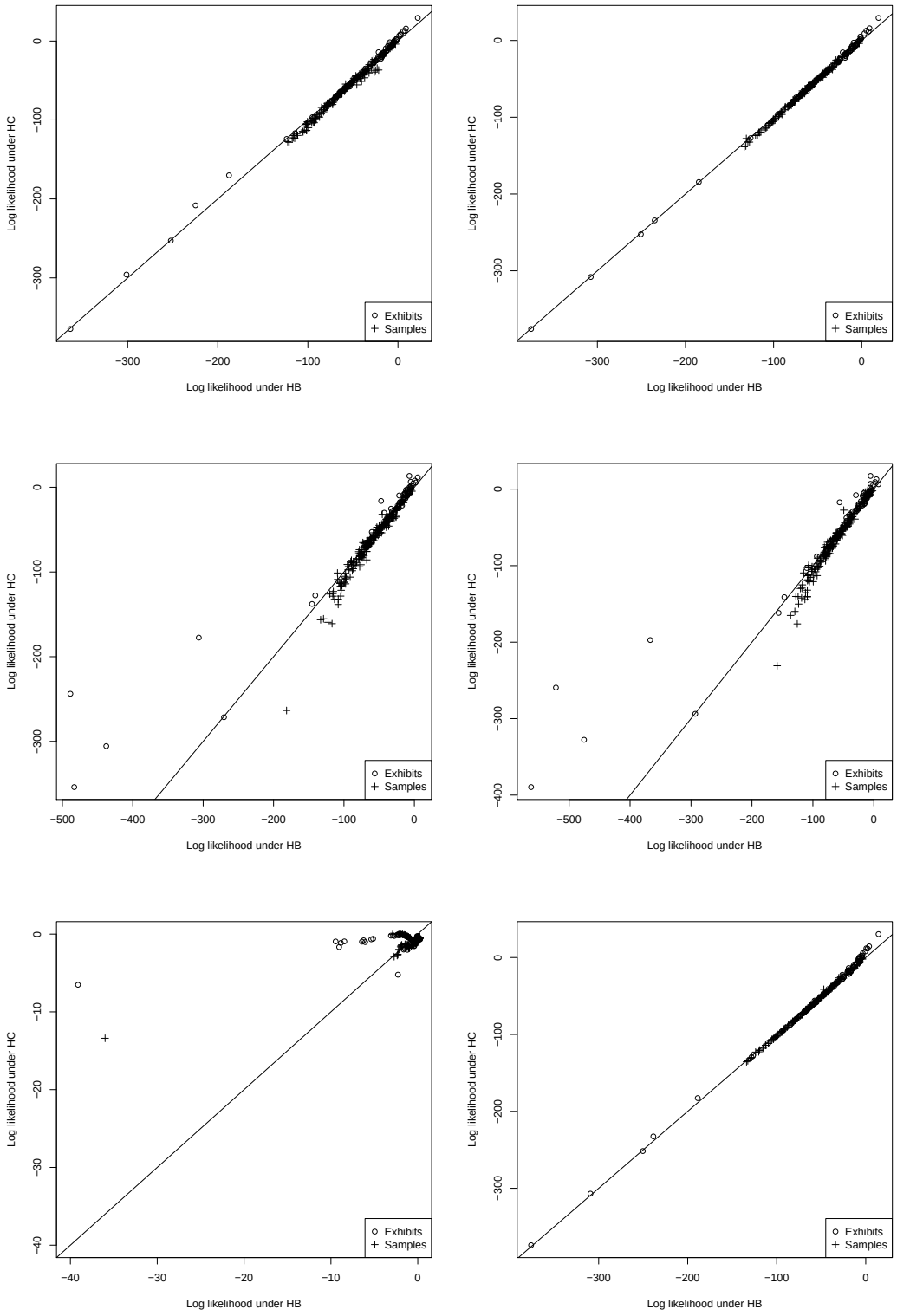

Figure 6.9: Scatter plots of likelihood values. Clockwise from top left - hidden Markov model, AR(1), adaptive bandwidth, AR(1) with random effects, standard model (assuming independence), fixed bandwidth

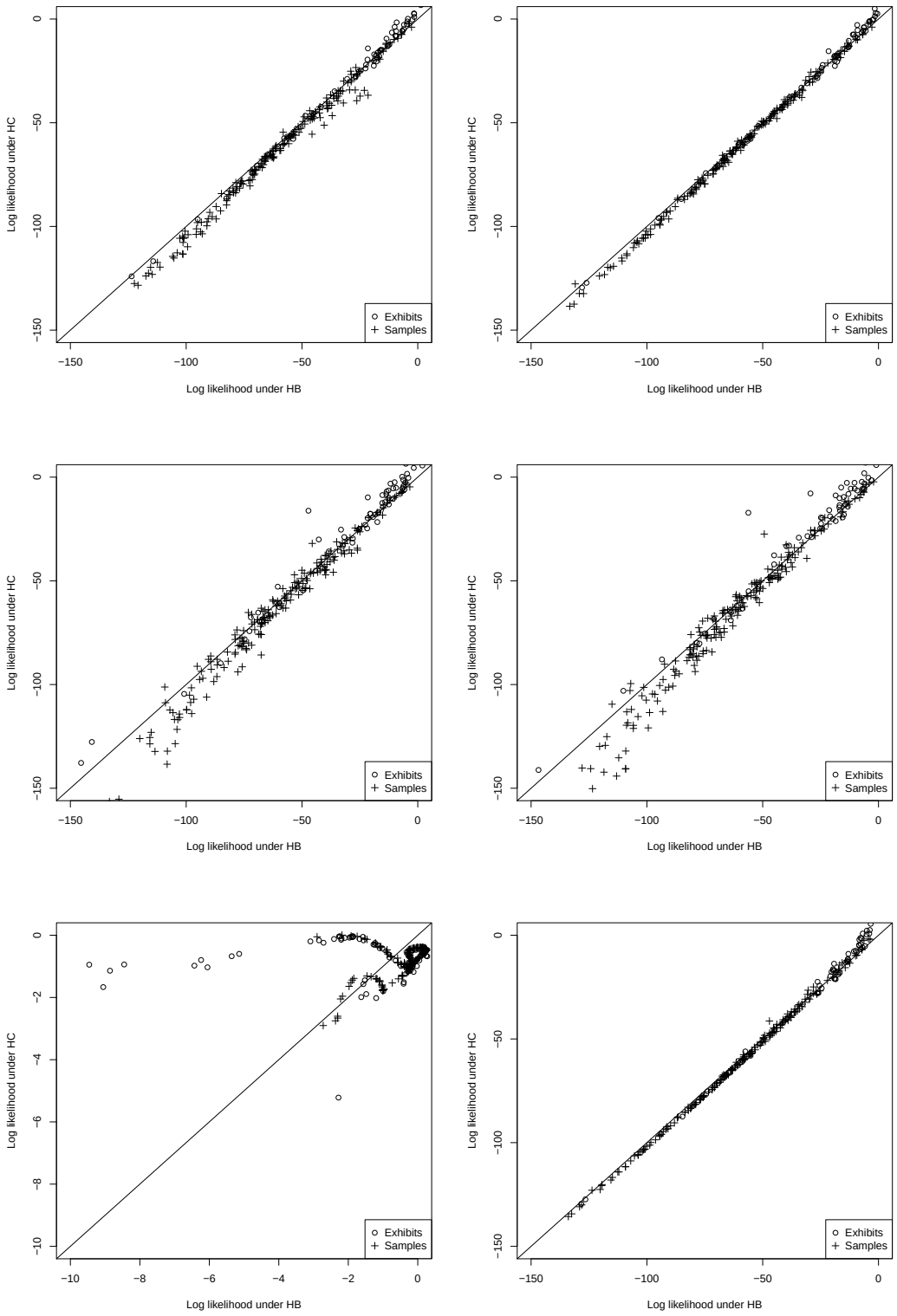

Figure 6.10: Scatter plots of likelihood values without outliers. Clockwise from top left - hidden Markov model, AR(1), adaptive bandwidth, AR(1) with random effects, standard model (assuming independence),fixed bandwidth

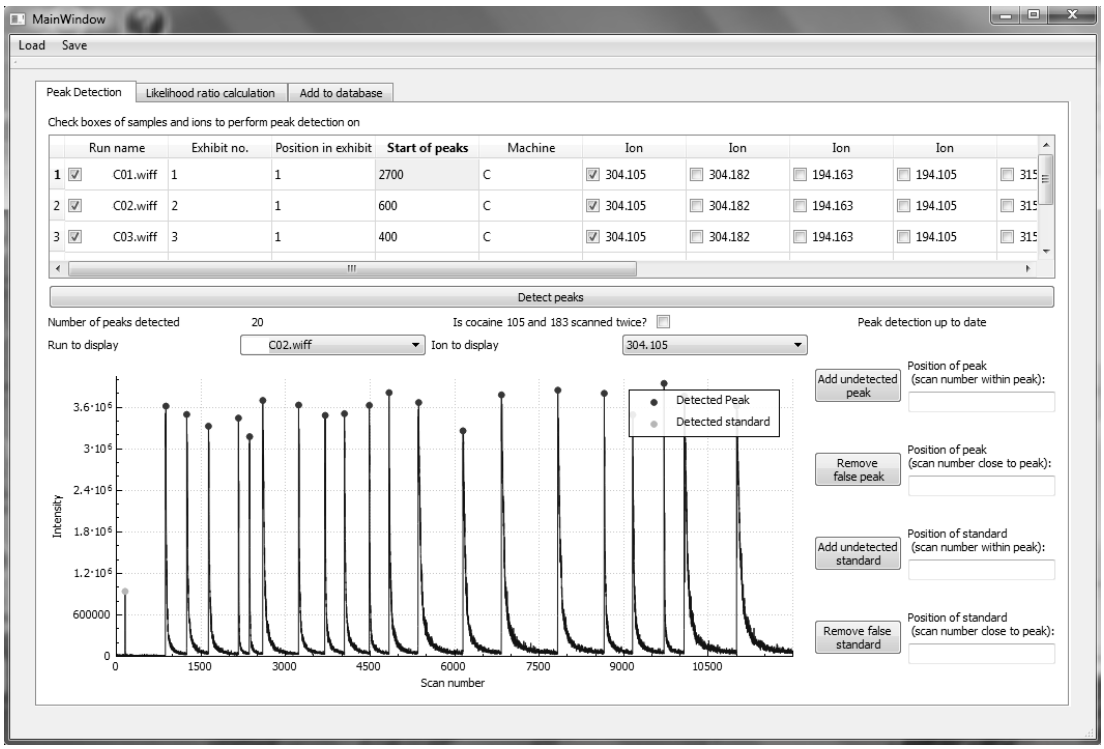

Figure 6.11: Peak detection with the GUI

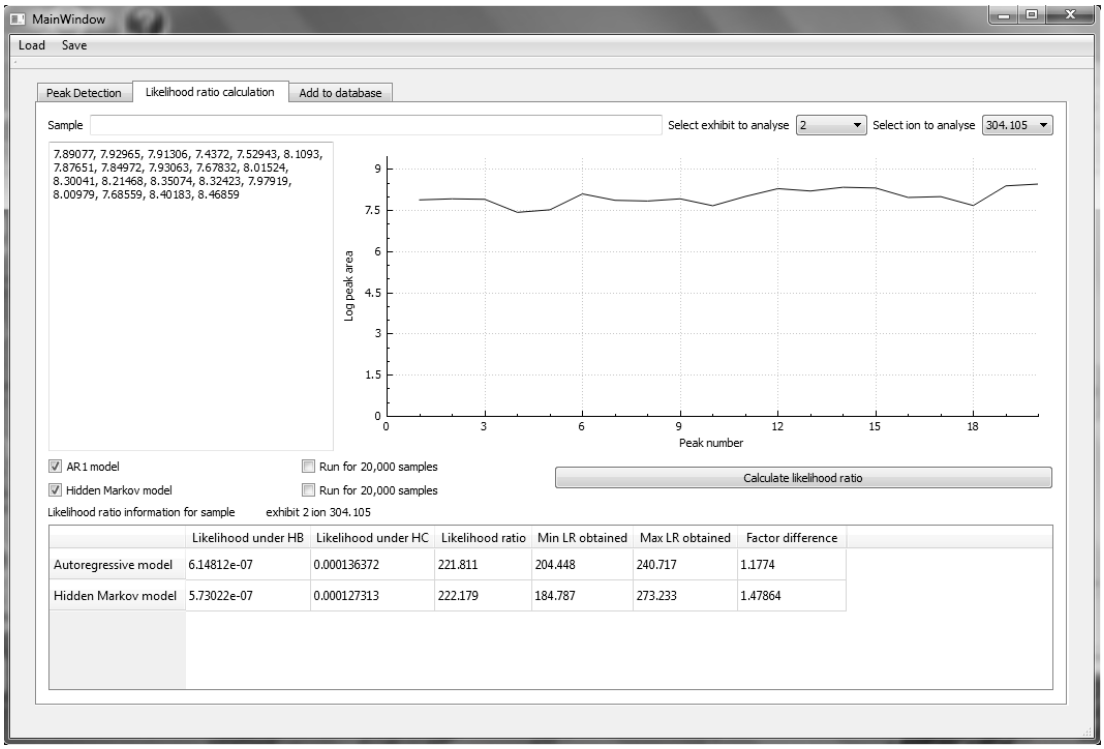

Figure 6.12: Likelihood ratio calculation with the GUI

Chapter 7

Conclusion

The two aims of this work were to develop methodology to evaluate likelihood ratios for autocorrelated evidential data, and to use this methodology to evaluate evidence relating to traces of cocaine on banknotes. The evidential data considered consisted of n univariate and autocorrelated measurements, given by $(z_1, \ldots, z_n)$. The problem considered was a discrimination problem; the likelihood ratio was required to evaluate the support given by the evidential data to propositions pertaining to which of two populations the evidence originated from. Previous work on the evaluation of likelihood ratios for continuous and multivariate evidential data has assumed that the multivariate measurements z_t for $t \in \{1, \ldots, n\}$ are independent of each other and that each z_t has a multivariate Normal distribution. Either the mean of the multivariate Normal distribution (Aitken and Lucy (2004)) or the mean and covariance matrix (Bozza et al. (2008)) were assumed to vary between samples according to a given between sample distribution. In this work, each measurement z_t was univariate, and the measurements $(z_1, \ldots, z_n)$ were assumed to be autocorrelated, rather than independent. A parameter α was introduced to model autocorrelation at lag one. The mean, variance and this autocorrelation parameter were assumed to vary between samples.

In Chapters 3 and 4, methodology was developed to evaluate the likelihood ratio for autocorrelated evidential data in three different scenarios. The first scenario assumed that the data followed an autoregressive process of lag one, with Normally distributed errors. The second scenario assumed that the data could be modelled using a hidden Markov model. This meant that, in addition to the assumptions of lag one autocorrelation and Normally distributed errors seen for the autoregressive model, the parameters determining the probability density function of each observation, conditional on the previous observation in the sample, were driven by an underlying latent Markov chain. The use of latent states allowed for the modelling of two different mean and variance levels within one sample of evidential data. Lastly, a nonparametric model was developed which accounts for autocorrelation at lag one and makes no assumption of Normality of the errors.

In Chapters 5 and 6 the evaluation of evidence specifically relating to traces of cocaine on banknotes was considered. In Chapter 5, a peak detection algorithm was developed to estimate the area

under each of a series of peaks, with each peak area giving a measure of the amount of cocaine on a banknote, as measured by a mass spectrometer. Two data sets were compiled from existing data, with the peak areas calculated using this peak detection algorithm. One data set consisted of exhibits of banknotes known to be associated with a person who was convicted of a crime involving cocaine (C) and one data set consisted of samples of banknotes from general circulation (B). Previous work on the evaluation of evidence for drug traces on banknotes (Besson (2004); Dixon et al. (2006); Jourdan et al. (2013)) has used a database of banknotes for C which have been seized by law enforcement agencies, meaning that banknotes in C are not necessarily associated with a person who is associated with crime. The definition used here for C is an improvement because the banknotes in the set C are known to have been associated with a suspect who was subsequently convicted of a crime involving cocaine. There is still work to be done, as the evidence relating to the cocaine traces on the banknotes was, in some cases, used as evidence to convict the suspect. This could result in biased data because an exhibit consisting of highly contaminated banknotes is more likely to have influenced the result of a trial. Another issue with the definition used for the set C is that even though the suspect was involved with criminal activity relating to cocaine, the banknotes themselves might not have been involved in the crime. Ideally a database of banknotes which were known to have been contaminated in the course of illegal activity involving cocaine would be constructed, without the evidence influencing the trial, although collection of such data may be impractical.

Limitations are also imposed by the database of general circulation banknotes. These banknotes form a convenience sample and were mainly obtained from banks in the Bristol area. There has been some work done to establish that the quantity of cocaine on banknotes does not vary between regions (Ebejer, Lloyd et al. (2007)), but it might still be necessary to tailor the database used to the region in which the crime occurred, or to use regional factors.

The evidential data consisted of the logarithms of the peak areas for each banknote in a sample seized by a law enforcement agency. Two propositions were considered for the development of likelihood ratios for this seized sample. These were H_C, that the seized sample is associated with a person who is involved with drug crime relating to cocaine, and H_B, that the seized sample is associated with a person who is not involved with drug crime relating to cocaine. The propositions being considered in practice may differ from these. For example, the suspect may maintain that he or she obtained the banknotes from the sale of a large item for cash. The two training data sets and the results discussed here relate only to the two propositions H_C and H_B. Any changes in either of the two propositions would require an associated change in the two databases, and it is not known whether the methods discussed here would still be effective.

Previous methods for the evaluation of evidence relating to traces of drugs on banknotes have considered the proportion of contaminated banknotes in a sample as a measure of the contamination (Ebejer, Brereton et al. (2005)). However, many banknotes in general circulation are contaminated with cocaine, so an approach based on such methods is not discriminatory for drug traces relating to cocaine. Another method (Besson (2004)) calculates the likelihood ratio for the intensity of cocaine

contamination on one seized banknote. Methods described in Taroni, Aitken et al. (2006) calculate the likelihood ratio for the quantities of cocaine contamination found on each of a sample of banknotes, but independence is assumed between the measurements on those banknotes. In Chapter 5 it was found that there was autocorrelation between measurements of the quantity of cocaine on banknotes within the same sample or exhibit, and that a model taking lag one autocorrelation into account would be appropriate for many of the samples and exhibits of banknotes in the training data sets. As such, the autoregressive model of lag one and the nonparametric model, which accounts for lag one autocorrelation, can be applied to data relating to traces of cocaine on banknotes. It was also found in Chapter 5 that samples and exhibits of banknotes were often arranged in bundles. Different bundles were thought to have different levels of contamination, perhaps because the person with whom they were associated had obtained the different bundles from different locations. The hidden Markov model, which can model the bundle from which each banknote originates using a latent state, can be used to model this situation. To fulfil the second aim of this work, these models were applied to the evidential data relating to traces of cocaine on banknotes.

In Chapter 6, the results obtained for each of the models were compared, and comparisons were made with a model that assumes independence. It was recommended that the hidden Markov model was used to evaluate evidence relating to traces of cocaine on banknotes. The hidden Markov model had the lowest rate of misleading evidence for samples of banknotes from general circulation: 10.4% of samples were found to provide misleading evidence when treated as the questioned sample. This rate of misleading evidence might still be considered to be large, but many of the 10.4% of general circulation samples had likelihood ratios close to one. Just four of the 193 general circulation samples had likelihood ratios which were larger than ten for the hidden Markov model (the largest was 44). The use of the hidden Markov model also resulted in likelihood ratios for some crime exhibits which were large enough to be used in practice to provide support for the proposition that the banknotes were associated with a person who was involved in drug crime relating to cocaine. Use of models which did not account for different bundles of banknotes within an exhibit having different levels of contamination was shown to result in smaller likelihood ratios for exhibits with both a large number of very highly contaminated banknotes and also a large number of banknotes with low contamination.

A subset of the exhibits in training data set C had similar levels of contamination to samples from general circulation in training data set B. It is thought that this is because some exhibits may not have been involved in criminal activity relating to cocaine (even though the person with whom they were associated was involved in such activity). In Section 5.3.3 it was shown that a well fitting model would assign a misleading likelihood ratio of less than one to a seized sample which was known to be in this subset. This means that low rates of misleading evidence cannot be achieved, even for well fitting models, for seized samples known to be from training data set C. As such, the rates of misleading evidence for crime exhibits were not used to assess the models.

It was found that the model assuming independence gave larger likelihood ratios to small exhibits than the models not assuming independence. This could imply that not taking autocorrelation into

account results in overstated likelihood ratios for small exhibits, which would mean that the model assuming independence assigns more support to H_C than is warranted by the data for these exhibits.

Two different methods for the evaluation of the likelihood ratio for the autoregressive model were compared. One (random effects) assumed Normal, inverse gamma and truncated Normal distributions for the between sample distributions of the mean, variance and autocorrelation parameters respectively. The other used a method which took the between sample distribution to be a weighted sum of the posterior distributions found for the mean, variance and autocorrelation parameters for each individual sample or exhibit in the training data set. The two methods had similar rates of misleading evidence for general circulation samples, but the method using a weighted sum of posterior distributions was found to be a better fit for seized samples of known origin.

The nonparametric methods were found to have large rates of misleading evidence, and many outliers. It is thought that this is because of a lack of robustness of these models where there are few training data. Increasing the amount of training data could therefore improve the performance of the nonparametric methods.

The methods developed here have provided a first step for the evaluation of likelihood ratios for continuous autocorrelated evidential data. These methods have been tested on data relating to traces of cocaine on banknotes, data which is known to be autocorrelated. The parametric models considered here were found to be effective at providing a statistically robust methodology for the evaluation of likelihood ratios for this type of evidence, without making the independence assumption seen in previous work, an assumption which can result in the overstating of likelihood ratios.

These methods evaluate likelihood ratios for univariate autocorrelated data. Future research can generalise this work to consider multivariate autocorrelated data. Data relating to drug traces on banknotes with twenty variables are available. These variables relate to five different drugs, each with two product ions. For each of the product ions, the peak area and the peak maximum for each banknote are available. This extra information might improve the discrimination between crime exhibits and general circulation samples.

The models discussed here can be extended to consider autocorrelation at larger lags. For the autoregressive model and the hidden Markov model, non-Normal errors can be considered. Likelihood ratios can also be evaluated for other forms of time series data such as those which follow an autoregressive-moving-average model. The hidden Markov model can also be generalised to consider both a larger number of states, and an unknown number of states. Currently, the number of states is assumed to be four (corresponding to two mean and variance levels). With more states, more levels of contamination could be modelled. As there are often many bundles of banknotes within each sample or exhibit, and each of these bundles may have a different origin, more levels of contamination may more accurately reflect the different levels of contamination on bundles within an exhibit or sample of banknotes. If the number of states is denoted by k, the generalisation corresponding to an unknown number of states would require the development of a method for estimating the posterior distribution of k, conditional on the data in each of the training data sets, B and C. This posterior distribution

could be estimated using a reversible jump Markov chain Monte Carlo algorithm. These extensions would increase the number of autocorrelated data structures to which likelihood ratio methods of evidence evaluation could be applied.

Glossary and notation

Glossary

B	Data set associated with proposition H_B, consisting of data $\mathbf{x}$. In chapters 5 and 6, data set of banknotes from general circulation.
Bundle	Banknotes within the same exhibit or sample are often grouped into separate bundles. These bundles may be fastened with elastic bands, bank strips, or arranged into 'dealer's wraps' with one banknote wrapped around a number of other banknotes.
Case	Multiple exhibits from the same criminal case make up a case. Law enforcement agencies choose how the banknotes gathered from one criminal case are divided into separate exhibits. Often this is based on the location in which the separate exhibits are found.
C	Data set associated with proposition H_C, consisting of data $\mathbf{y}$. In chapters 5 and 6, data set of banknotes that were involved in a criminal case in which the suspect was convicted of a crime involving cocaine.
D	General set of banknotes. Replace D by B if proposition H_B is being considered and replace D by C if proposition H_C is being considered. Consists of general data $\mathbf{w}$.
E	Used to denote evidence, which is usually in the form of a set of measurements.
Exhibit	Used in chapters 5 and 6 to denote a number of banknotes which were obtained from law enforcement agencies in one group. Exhibits are known to be associated with a person who was involved with drug crime relating to cocaine. Multiple exhibits make up the set C. Measurements on the i-th exhibit in set C are denoted $\mathbf{y}_i$.
H_B	Chapters 3 and 4: proposition that the evidential data are from set B. Chapters 5 and 6: proposition that the banknotes are associated with a person who is not involved in criminal activity involving cocaine.
H_C	Chapters 3 and 4: proposition that the evidential data are from set C.

	Chapters 5 and 6: proposition that the banknotes are associated with a person who is involved in criminal activity involving cocaine.
H_D	General proposition, should replace with H_C or H_B as appropriate.
H_d	Defence proposition.
H_p	Prosecution proposition.
M	Parameter denoting the model choice. Used with subscript i to represent model choice for i-th sample or exhibit. Used with subscript z to represent model choice for seized sample $\mathbf{z}$.
M_A	Parameter representing the choice of the autoregressive model.
M_{A_r}	Parameter representing the choice of the autoregressive model with random effects.
M_H	Parameter representing the choice of the hidden Markov model.
m_B	Number of samples in set B.
m_C	Number of samples (chapters 3 and 4) or exhibits (chapters 5 and 6) in set C.
n	Number of observations in seized sample $\mathbf{z}$.
n_{B_i}	Number of observations in i-th sample from set B, $\mathbf{x}_i$.
n_{C_i}	Number of observations in i-th sample from set C, $\mathbf{y}_i$.
N	Number of Metropolis-Hastings draws.
Peak area	The area under one peak (representing one banknote) of the mass spectrometer output, as calculated using the peak area algorithm described in section 5.4. Gives a measure of the quantity of cocaine on a banknote.
Run	The mass spectrometers used to analyse samples and exhibits of banknotes are set up to analyse drug traces for a period of twenty minutes. Any sample or exhibit which takes longer than this to analyse must be analysed in separate runs. A run is a single twenty minute period of analysis. One exhibit or sample can consist of multiple runs.
Sample	Used in chapters 5 and 6 to mean a number of banknotes from general circulation that were found in one location (or obtained in one transaction). Multiple samples make up the set B. Measurements on the i-th sample in set B are denoted $\mathbf{x}_i$. Used elsewhere to mean a number of observations representing one item in either set B or set C. Multiple samples make up one set. Measurements on the i-th sample in set B are denoted $\mathbf{x}_i$. Measurements on the i-th sample in set C are denoted $\mathbf{y}_i$.

Seized sample	A sample of evidential data for which the likelihood ratio is to be evaluated. Measurements from the seized sample are denoted by **z**. It is unknown which data set a seized sample belongs to. Also known as the questioned sample.
Set	Used to refer to an entire data set, either B, C or the general set D.
v_i	Weight for the i-th sample or exhibit. Used to obtain the between sample distribution as a weighted sum of the posterior distributions of the parameters associated with each individual sample and exhibit in chapter 4.
w	General training set associated with general proposition H_D. Replace **w** with **x** or **y** as appropriate.
$\mathbf{w}_i = (w_{i1}, \ldots, w_{in_{D_i}})$	i-th sample from **w**.
w_{it}	t-th observation from i-th sample in **w**.
x	Training data set associated with proposition H_B.
$\mathbf{x}_i = (x_{i1}, \ldots, x_{in_{B_i}})$	i-th sample from **x**.
x_{it}	t-th observation from i-th sample in **x**.
y	Training data set associated with proposition H_C.
$\mathbf{y}_i = (y_{i1}, \ldots, y_{in_{C_i}})$	i-th sample from **y**.
y_{it}	t-th observation from i-th sample in **y**.
$\mathbf{z} = (z_1, \ldots, z_n)$	Seized sample, also known as questioned sample.

Model parameters

Model parameters are used with subscript A (autoregressive model), A_r (autoregressive model with random effects) or H (hidden Markov model) to represent the parameters for a specific model. They are used with subscript i to represent parameters for the i-th sample or exhibit and they are used with superscript C (associated with proposition H_C) or B (associated with proposition H_B) to represent the parameters for a specific training data set.

θ	Vector of model parameters. Can be used to represent the model parameters of the autoregressive model (with or without random effects) or the hidden Markov model. The parameter θ is used with a subscript when it is needed to denote the specific model being represented.
μ	Mean of autoregressive models.
μ_1	Mean associated with the current banknote for states 1 and 3 of hidden Markov model. Is the smaller mean so that $\mu_1 < \mu_2$.

μ_2	Mean associated with the current banknote for states 2 and 4 of hidden Markov model. Is the larger mean so that $\mu_2 > \mu_1$.
σ^2	Variance of error terms for autoregressive models.
σ_1^2	Variance of error terms for observations in states 1 and 3 of hidden Markov model.
σ_2^2	Variance of error terms for observations in states 2 and 4 of hidden Markov model.
β	Hyperparameter of σ^2 for autoregressive model without random effects.
β_1	Hyperparameter of σ_1^2.
β_2	Hyperparameter of σ_2^2.
α	Autocorrelation parameter for autoregressive models.
α_1	Autocorrelation parameter for hidden Markov model.
p_{01}	Hidden Markov model transition probability, probability of moving from state 1 to state 2 and of moving from state 3 to state 2.
p_{10}	Hidden Markov model transition probability, probability of moving from state 2 to state 3 and of moving from state 4 to state 3.
p_1, p_2, p_3, p_4	Probability that the first observation in a sample or exhibit is in state 1 (p_1),2 (p_2),3 (p_3) and 4 (p_4) when using the hidden Markov model.
$\mathbf{S}_B = \{S_{B_{it}}; i = 1,\ldots,m_B, t = 1,\ldots,n_{B_i}\}$	Used with the hidden Markov model to represent the hidden states associated with data $\mathbf{x}$. A state $S_{B_{it}}$ is associated with each observation x_{it}. Each state $S_{B_{it}}$ can take values in $\{1,2,3,4\}$.
$\mathbf{S}_C = \{S_{C_{it}}; i = 1,\ldots,m_C, t = 1,\ldots,n_{C_i}\}$	Used with the hidden Markov model to represent the hidden states associated with data $\mathbf{y}$. A state $S_{C_{it}}$ is associated with each observation y_{it}. Each state $S_{C_{it}}$ can take values in $\{1,2,3,4\}$.
$\mathbf{R} = \{R_t; t = 1,\ldots,n\}$	Used with the hidden Markov model to represent the hidden states associated with seized sample $\mathbf{z}$ with one state for each of the n observations. Each state R_t can take values in $\{1,2,3,4\}$.

Bibliography

Ailliot, P. and V. Monbet (2012). 'Markov-switching autoregressive models for wind time series'. *Environmental Modelling & Software* 30, pp. 92–101.

Aitken, C. G. G. and E. Gold (2013). 'Evidence evaluation for discrete data'. *Forensic Science International* 230.1-3, pp. 147–155.

Aitken, C. G. G. and D. Lucy (2004). 'Evaluation of trace evidence in the form of multivariate data'. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 53.1, pp. 109–122.

Aitken, C. G. G., D. Lucy, G. Zadora and J. M. Curran (2006). 'Evaluation of transfer evidence for three-level multivariate data with the use of graphical models'. *Computational Statistics & Data Analysis* 50.10, pp. 2571–2588.

Aitken, C. G. G., P. Roberts and G. Jackson (2010). *Fundmentals of probability and statistical evidence in criminal proceedings*. Available on http://www.rss.org.uk/statsandlaw, last accessed 7th November 2013. London: The Royal Statistical Society.

Aitken, C. G. G., Q. Shen, R. Jensen and B. Hayes (2007). 'The evaluation of evidence for exponentially distributed data'. *Computational Statistics & Data Analysis* 51.12, pp. 5682–5693.

Aitken, C. G. G. and D. Stoney (1991). *The Use of Statistics in Forensic Science*. Chichester: Ellis Horwood Limited.

Aitken, C. G. G. and F. Taroni (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. 2nd ed. Chichester: Wiley.

Aitken, C. G. G., G. Zadora and D. Lucy (2007). 'A two-level model for evidence evaluation'. *Journal of Forensic Sciences* 52.2, pp. 412–419.

Alberink, I., A. Bolck and S. Menges (2013). 'Posterior likelihood ratios for evaluation of forensic trace evidence given a two-level model on the data'. *Journal of Applied Statistics* 40.12, pp. 2579–2600.

Albert, J. H. and S. Chib (1993). 'Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts'. *Journal of Business & Economic Statistics* 11.1, pp. 1–15.

Armenta, S. and M. de la Guardia (2008). 'Analytical methods to determine cocaine contamination of banknotes from around the world'. *TrAC Trends in Analytical Chemistry* 27.4, pp. 344–351.

Bengtsson, H. (2013). *R.matlab: Read and write of MAT files together with R-to-Matlab connectivity*. R package version 1.7.0.

Besson, L. (2004). 'Détection des stupéfiants par IMS'. Master's thesis. Universite de Lausanne.

Bozza, S., F. Taroni, R. Marquis and M. Schmittbuhl (2008). 'Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship'. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 57.3, pp. 329–341.

Breiman, L., W. Meisel and E. Purcell (1977). 'Variable kernel estimates of multivariate densities'. *Technometrics* 19.2, pp. 135–144.

Brereton, R. G. (2009). *Chemometrics for Pattern Recognition*. Chichester: Wiley.

Brooks, S. P. and A. Gelman (1998). 'General methods for monitoring convergence of iterative simulations'. *Journal of Computational and Graphical Statistics* 7.4, pp. 434–455.

Buckleton, J. S., C. M. Triggs and S. J. Walsh (2005). *Forensic DNA Evidence Interpretation*. Boca Raton: CRC press.

Cappé, O., E. Moulines and T. Rydén (2005). *Inference in Hidden Markov Models*. Springer Series in Statistics. New York: Springer.

Carter, J. F., R. Sleeman and J. Parry (2003). 'The distribution of controlled drugs on banknotes via counting machines'. *Forensic Science International* 132.2, pp. 106–112.

Chatfield, C. (2004). *The Analysis of Time Series: An Introduction*. 6th ed. London: Chapman and Hall.

Chib, S. (1995). 'Marginal likelihood from the Gibbs output'. *Journal of the American Statistical Association* 90.432, pp. 1313–1321.

— (1996). 'Calculating posterior distributions and modal estimates in Markov mixture models'. *Journal of Econometrics* 75.1, pp. 79–97.

Chib, S. and E. Greenberg (1995). 'Understanding the Metropolis-Hastings algorithm'. *The American Statistician* 49.4, pp. 327–335.

Chib, S. and I. Jeliazkov (2001). 'Marginal likelihood from the Metropolis-Hastings output'. *Journal of the American Statistical Association* 96.453, pp. 270–281.

Cook, R., I. W. Evett, G. Jackson, P. J. Jones and J. A. Lambert (1998a). 'A hierarchy of propositions: deciding which level to address in casework'. *Science & Justice* 38.4, pp. 231–239.

— (1998b). 'A model for case assessment and interpretation'. *Science & Justice* 38.3, pp. 151–156.

Davis, L. J., C. P. Saunders, A. Hepler and J. Buscaglia (2012). 'Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios'. *Forensic Science International* 216.1-3, pp. 146–157.

Dixon, S. J., R. G. Brereton, J. F. Carter and R. Sleeman (2006). 'Determination of cocaine contamination on banknotes using tandem mass spectrometry and pattern recognition'. *Analytica Chimica Acta* 559.1, pp. 54–63.

Du, P., W. A. Kibbe and S. M. Lin (2006). 'Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching'. *Bioinformatics* 22.17, pp. 2059–2065.

Ebejer, K. A., R. G. Brereton, J. F. Carter, S. L. Ollerton and R. Sleeman (2005). 'Rapid comparison of diacetylmorphine on banknotes by tandem mass spectrometry'. *Rapid Communications in Mass Spectrometry* 19.15, pp. 2137–2143.

Ebejer, K. A., G. R. Lloyd, R. G. Brereton, J. F. Carter and R. Sleeman (2007). 'Factors influencing the contamination of UK banknotes with drugs of abuse'. *Forensic Science International* 171.2-3, pp. 165–170.

Ebejer, K. A., J. Winn, J. F. Carter, R. Sleeman, J. Parker and F. Körber (2007). 'The difference between drug money and a "lifetime's savings"'. *Forensic Science International* 167.2-3, pp. 94–101.

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. New York: Springer.

Eddelbuettel, D. and R. Francois (2011). 'Rcpp: seamless R and C++ integration'. *Journal of Statistical Software* 40.8, pp. 1–18.

Evett, I. W., G. Jackson, J. A. Lambert and S. McCrossan (2000). 'The impact of the principles of evidence interpretation on the structure and content of statements.' *Science & Justice* 40.4, pp. 233–239.

Evett, I. W., R. Pinchin and C. Buffery (1992). 'An investigation of the feasibility of inferring ethnic origin from DNA profiles'. *Journal of the Forensic Science Society* 32.4, pp. 301–306.

Evett, I. W. (1983). 'What is the probability that this blood came from that person? A meaningful question?' *Journal of the Forensic Science Society* 23.1, pp. 35–39.

Evett, I. W., G. Jackson and J. A. Lambert (2000). 'More on the hierarchy of propositions: exploring the distinction between explanations and propositions'. *Science & Justice* 40.1, pp. 3–10.

Fan, J., Q. Yao and H. Tong (1996). 'Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems'. *Biometrika* 83.1, pp. 189–206.

Frühwirth-Schnatter, S. (2001). 'Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models'. *Journal of the American Statistical Association* 96.453, pp. 194–209.

— (2004). 'Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques'. *Econometrics Journal* 7.1, pp. 143–167.

— (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. New York: Springer.

Gassiat, E. and J. Rousseau (2013). 'On the asymptotic behaviour of the posterior distribution in hidden Markov models'. *To appear in Bernoulli*, available on http://www.math.u-psud.fr/ gassiat/HMM.pdf, last accessed 6th November 2013.

Gelfand, A. E., A. F. M. Smith and T.-M. Lee (1992). 'Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling'. *Journal of the American Statistical Association* 87.418, pp. 523–532.

Gelman, A. (2006). 'Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)'. *Bayesian Analysis* 1.3, pp. 515–534.

Gelman, A., G. Roberts and W. Gilks (1996). 'Efficient Metropolis jumping rules'. *Bayesian Statistics* 5, pp. 599–608.

Gelman, A. and D. B. Rubin (1992). 'Inference from iterative simulation using multiple sequences'. *Statistical Science* 7.4, pp. 457–472.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin (2004). *Bayesian Data Analysis*. 2nd ed. London: Chapman and Hall.

Geman, S. and D. Geman (1984). 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images'. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 6.6, pp. 721–741.

Gonzalez-Rodriguez, J., A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar and J. Ortega-Garcia (2006). 'Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition'. *Computer Speech & Language* 20.2-3, pp. 331–355.

Green, P. J. (1995). 'Reversible jump Markov chain Monte Carlo computation and Bayesian model determination'. *Biometrika* 82.4, pp. 711–732.

Green, P. J. (2003). 'Trans-dimensional Markov chain Monte Carlo'. In: *Highly Structured Stochastic Systems*. Ed. by P. J. Green, N. Hjort and S. Richardson. Oxford: Oxford University Press.

Hall, P., J. Racine and Q. Li (2004). 'Cross-validation and the estimation of conditional probability densities'. *Journal of the American Statistical Association* 99.468, pp. 1015–1026.

Hamilton, J. D. (1989). 'A new approach to the economic analysis of nonstationary time series and the business cycle'. *Econometrica: Journal of the Econometric Society* 57.2, pp. 357–384.

Hastings, W. K. (1970). 'Monte Carlo sampling methods using Markov chains and their applications'. *Biometrika* 57.1, pp. 97–109.

Hayfield, T. and J. S. Racine (2008). 'Nonparametric econometrics: the np package'. *Journal of Statistical Software* 27.5, pp. 1–32.

Jenkins, A. J. (2001). 'Drug contamination of US paper currency'. *Forensic Science International* 121.3, pp. 189–193.

Jourdan, T. H., A. M. Veitenheimer, C. K. Murray and J. R. Wagner (2013). 'The quantitation of cocaine on US currency: survey and significance of the levels of contamination'. *Journal of Forensic Sciences* 58.3, pp. 616–624.

Kim, C.-J. and C. R. Nelson (1999). 'Has the US economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle'. *Review of Economics and Statistics* 81.4, pp. 608–616.

Lindley, D. V. (1977). 'A problem in forensic science'. *Biometrika* 64.2, pp. 207–213.

Liu, J. S. (1994). 'The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem'. *Journal of the American Statistical Association* 89.427, pp. 958–966.

Lloyd, G. (2009). 'Chemometrics and pattern recognition for the analysis of multivariate datasets'. Doctoral dissertation. University of Bristol.

Lu, H.-M., D. Zeng and H. Chen (2010). 'Prospective infectious disease outbreak detection using Markov switching models'. *Knowledge and Data Engineering, IEEE Transactions on* 22.4, pp. 565–577.

McCulloch, R. E. and P. E. Rossi (1992). 'Bayes factors for nonlinear hypotheses and likelihood distributions'. *Biometrika* 79.4, pp. 663–676.

McCulloch, R. E. and R. S. Tsay (1994). 'Statistical analysis of economic time series via Markov switching models'. *Journal of Time Series Analysis* 15.5, pp. 523–539.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller (1953). 'Equation of state calculations by fast computing machines'. *The Journal of Chemical Physics* 21, pp. 1087–1092.

Morrison, G. S. (2011). 'Measuring the validity and reliability of forensic likelihood-ratio systems'. *Science & Justice* 51.3, pp. 91–98.

Neumann, C., I. W. Evett and J. Skerrett (2012). 'Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm'. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175.2, pp. 371–415.

Neumann, C., C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz and A. Bromage-Griffiths (2007). 'Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae'. *Journal of Forensic Sciences* 52.1, pp. 54–64.

Newton, M. A. and A. E. Raftery (1994). 'Approximate Bayesian inference with the weighted likelihood bootstrap'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 56.1, pp. 3–48.

Pierce, D. (2011). *ncdf: Interface to Unidata netCDF data files*. R package version 1.6.6.

Plummer, M. (2013). *rjags: Bayesian graphical models using MCMC*. R package version 3-10.

Plummer, M., N. Best, K. Cowles and K. Vines (2006). 'CODA: convergence diagnosis and output analysis for MCMC'. *R News* 6.1, pp. 7–11.

Rabiner, L. R. (1989). 'A tutorial on hidden Markov models and selected applications in speech recognition'. *Proceedings of the IEEE* 77.2, pp. 257–286.

Ramsey, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. 2nd ed. Springer Series in Statistics. New York: Springer.

Richardson, S. and P. J. Green (1997). 'On Bayesian analysis of mixtures with an unknown number of components (with discussion)'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.4, pp. 731–792.

Robert, C. P. (1995). 'Simulation of truncated Normal variables'. *Statistics and Computing* 5.2 (2), pp. 121–125.

Robert, C. P. and G. Casella (2004). *Monte Carlo Statistical Methods*. New York: Springer.

Roberts, D. J., J. F. Carter, R. Sleeman and I. F. A. Burton (1997). 'Application of tandem mass spectrometry to the detection of drugs on cash'. *Spectroscopy Europe* 9.6, pp. 24–27.

Rydén, T. (2008). 'EM versus Markov chain Monte Carlo for estimation of hidden Markov models: a computational perspective'. *Bayesian Analysis* 3.4, pp. 659–688.

Scott, S. L. (2002). 'Bayesian methods for hidden Markov models'. *Journal of the American Statistical Association* 97.457, pp. 337–351.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Sleeman, R., I. F. A. Burton, J. F. Carter, D. Roberts and P. Hulmston (2000). 'Drugs on money'. *Analytical Chemistry* 72.11, 397A–403A.

Snijders, T. A. and R. J. Bosker (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. London: SAGE.

Spezia, L. (2010). 'Bayesian analysis of multivariate Gaussian hidden Markov models with an unknown number of regimes'. *Journal of Time Series Analysis* 31.1, pp. 1–11.

Taroni, F., C. G. G. Aitken, P. Garbolino and A. Biedermann (2006). *Bayesian Networks and Probabilistic Inference in Forensic Science*. Chichester: Wiley.

Taroni, F., S. Bozza, A. Biedermann, P. Garbolino and C. G. G. Aitken (2010). *Data Analysis in Forensic Science: a Bayesian Decision Perspective*. Chichester: Wiley.

Terrell, G. R. and D. W. Scott (1992). 'Variable kernel density estimation'. *The Annals of Statistics* 20.3, pp. 1236–1265.

Wilson, A., C. G. G. Aitken, R. Sleeman and J. F. Carter (2013). 'The evaluation of evidence for autocorrelated data in relation to traces of cocaine on banknotes'. *In press, Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

— (2014). 'The evaluation of evidence relating to traces of cocaine on banknotes'. *Forensic Science International* 236, pp. 67–76.

Zadora, G., T. Neocleous and C. G. G. Aitken (2010). 'A two-level model for evidence evaluation in the presence of zeros'. *Journal of Forensic Sciences* 55.2, pp. 371–384.

Appendix A

Conditional distributions for Gibbs sampler

The prior distributions chosen for the parameters of the autoregressive model without random effects and the hidden Markov model are such that it is possible to obtain analytically the distributions of each of the parameters, conditional on the data and all of the other parameters. A Gibbs sampler can then be used to obtain draws from the joint distribution of the parameters, using these conditional distributions. A Gibbs sampler is a Markov chain Monte Carlo algorithm. It can be used for situations where it is wanted to obtain draws from the distribution of a d-dimensional parameter θ, conditional on a set of data $\mathbf{w}_i$, and where there are analytical expressions for all of the conditonal density functions $f(\theta_k \mid \mathbf{w}_i, \theta_{-k})$ for $k \in \{1,\ldots,d\}$ (where the notation θ_{-k} denotes the vector θ without the kth entry and θ_k is the k-th entry of the vector θ). Let the r-th draw of θ be denoted $\theta^{(r)}$. A Gibbs sampler would draw $\theta_1^{(r+1)}$ from $f(\theta_1 \mid \mathbf{w}_i, \theta_{-1}^{(r)})$, followed by $\theta_2^{(r+1)}$ from $f(\theta_2 \mid \mathbf{w}_i, \theta_1^{(r+1)}, \theta_{-1,-2}^{(r)})$ and so on. After a burn in period, these draws converge to the posterior distribution of θ conditional on the data $\mathbf{w}_i$, given in (3.3)(Geman and Geman (1984)). For more information and examples using the Gibbs sampler, see p287 of Gelman, Carlin et al. (2004).

Whilst the Gibbs sampler was not the preferred method of obtaining posterior draws for the example used in Chapters 5 and 6, it may be suitable for other data. In particular, using a Gibbs sampler avoids the potentially difficult and time consuming task of adjusting the proposal distributions in the Metropolis-Hastings algorithms discussed in Chapter 3, because the proposal distributions are simply the conditional distributions given in this appendix (the Gibbs sampler is a special case of the Metropolis-Hastings sampler). In this appendix, the conditional distributions for the Gibbs sampler are listed for the autoregressive model and the hidden Markov model. To obtain draws from the joint posterior distribution, each of these conditional distributions should be sampled from in turn, as described above. General notation is used, as in the main text, with $\mathbf{w}_i$ representing the i-th sample and D representing the proposition being considered

A.1 Autoregressive model

In this section, the distribution of each of the parameters $(\mu, \sigma^2, \beta, \alpha)$, conditional on a set of autocorrelated data $\mathbf{w}_i = (w_{i1}, \ldots, w_{in_{D_i}})$ and each of the remaining parameters is given. These distributions assume that the autoregressive model with lag one specified in Section 3.1 is being used to model the data $\mathbf{w}_i$. The likelihood $L(\mu, \sigma^2, \beta, \alpha)$ is given by

$$\begin{aligned} L(\mu, \sigma^2, \beta, \alpha) \propto (2\pi\sigma^2)^{-\frac{n_{D_i}}{2}} \exp\left[-\frac{1}{2\sigma^2}(w_{i1}-\mu)^2\right] \\ \times \exp\left[-\sum_{t=2}^{n_{D_i}}\left(\frac{1}{2\sigma^2}(w_{it}-\mu+\alpha\mu-\alpha w_{i,t-1})^2\right.\right]. \end{aligned} \tag{A.1}$$

A.1.1 μ

The prior distribution of μ was specified in Section 3.1.1 as a Normal distribution so that

$$\mu \sim N(\mu_0, V_\mu).$$

Combining this prior distribution with the likelihood (A.1) and discarding the constant terms, the conditional density function of μ is given by

$$\begin{aligned} f(\mu \mid \mathbf{w}_i, \sigma^2, \alpha, \beta) \propto \exp\left[-\frac{1}{2V_\mu}(\mu-\mu_0)^2 - \frac{1}{2\sigma^2}(w_{i1}-\mu)^2\right. \\ \left.-\sum_{t=2}^{n_{D_i}}\left(\frac{1}{2\sigma^2}(w_{it}-\mu+\alpha\mu-\alpha w_{i,t-1})^2\right)\right]. \end{aligned} \tag{A.2}$$

Letting

$$A_\mu = \left[\frac{1}{V_\mu} + \frac{1}{\sigma^2} + \frac{1}{\sigma^2}(n_{D_i}-1)(\alpha-1)^2\right]^{-1}$$

and

$$M_\mu = A_\mu\left[\frac{\mu_0}{V_\mu} + \frac{1}{\sigma^2}w_{i1} + \frac{1}{\sigma^2}\sum_{t=2}^{n_{D_i}}\left[(\alpha-1)(\alpha w_{i,t-1} - w_{it})\right]\right],$$

equation (A.2) becomes

$$f(\mu \mid \mathbf{w}_i, \sigma^2, \alpha, \beta) \propto \exp\left[-\frac{1}{2A_\mu}(\mu - M_\mu)^2\right]$$

and so the conditional distribution of μ is given by

$$\mu \mid \mathbf{w}_i, \sigma^2, \alpha, \beta \sim N(M_\mu, A_\mu).$$

A.1.2 σ^2

The prior distribution of σ^2 given the hyperparameter β is inverse gamma, with parameters

$$\sigma^2 \mid \beta \sim IG(\gamma, \beta),$$

so that,

$$f(\sigma^2 \mid \beta) \propto \beta^{\gamma} \sigma^{-(2\gamma+2)} \exp\left[-\frac{\beta}{\sigma^2}\right].$$

Combining the likelihood (A.1) with this prior density function, and discarding constant terms, the conditional density function for σ^2 is given by

$$f(\sigma^2 \mid \mathbf{w}_i, \mu, \alpha, \beta) \propto \sigma^{-(2\gamma+2)} \exp\left[-\frac{\beta}{\sigma^2}\right] (2\pi\sigma^2)^{-\frac{n_{D_i}}{2}} \exp\left[-\frac{1}{2\sigma^2}(w_{i1}-\mu)^2 - \sum_{t=2}^{n_{D_i}} \left(\frac{1}{2\sigma^2}(w_{it}-\mu+\alpha\mu-\alpha w_{i,t-1})^2\right)\right],$$

which simplifies to

$$f(\sigma^2 \mid \mathbf{w}_i, \mu, \alpha, \beta) \propto \sigma^{-(2\gamma+2+n_{D_i})} \exp\left[-\frac{1}{2\sigma^2}\left(2\beta + (w_{i1}-\mu)^2 + \sum_{t=2}^{n_{D_i}} (w_{it}-\mu+\alpha\mu-\alpha w_{i,t-1})^2\right)\right].$$

So, letting

$$D = \frac{1}{2}\left[2\beta + (w_{i1}-\mu)^2 + \sum_{t=2}^{n_{D_i}} (w_{it}-\mu+\alpha\mu-\alpha w_{i,t-1})^2\right],$$

the conditional distribution for σ^2 is given by

$$\sigma^2 \mid \mathbf{w}_i, \mu, \alpha, \beta \sim IG\left(\frac{2\gamma + n_{D_i}}{2}, D\right).$$

A.1.3 α

The prior distribution of α is Normal, restricted to lie between -1 and 1, so that

$$\alpha \sim N(\alpha_0, V_\alpha) I(|\alpha| < 1),$$

where $I(|\alpha| < 1)$ is the indicator function such that

$$I(|\alpha| < 1) = 1 \text{ if } |\alpha| < 1,$$
$$= 0 \text{ if } |\alpha| \geq 1.$$

Combining this with the likelihood (A.1), the conditional density function of α is proportional to

$$\begin{aligned} f(\alpha \mid \mathbf{w}_i, \mu, \sigma^2, \beta) \propto \exp\Bigg[& -\frac{1}{2V_\alpha}(\alpha - \alpha_0)^2 \\ & - \sum_{t=2}^{n_{D_i}} \left(\frac{1}{2\sigma^2} (w_{it} - \mu + \alpha\mu - \alpha w_{i,t-1})^2 \right) \Bigg] I(|\alpha| < 1) \\ \propto \exp\Bigg[& -\frac{1}{2} \left(\frac{1}{V_\alpha} + \sum_{t=2}^{n_{D_i}} \frac{1}{\sigma^2} (w_{i,t-1} - \mu)^2 \right) \times \\ & \left(\alpha - \frac{\frac{\alpha_0}{V_\alpha} + \sum_{t=2}^{n_{D_i}} \frac{1}{\sigma^2} (w_{i,t-1} - \mu)(w_{it} - \mu)}{\frac{1}{V_\alpha} + \sum_{t=2}^{n_{D_i}} \frac{1}{\sigma^2} (w_{i,t-1} - \mu)^2} \right)^2 \Bigg] I(|\alpha| < 1). \end{aligned}$$

So, letting

$$\begin{aligned} D_\alpha &= \left(\frac{1}{V_\alpha} + \sum_{t=2}^{n_{D_i}} \frac{1}{\sigma^2} (w_{i,t-1} - \mu)^2 \right)^{-1} \\ M_\alpha &= D_\alpha \left(\frac{\alpha_0}{V_\alpha} + \sum_{t=2}^{n_{D_i}} \frac{1}{\sigma^2} (w_{i,t-1} - \mu)(w_{it} - \mu) \right), \end{aligned}$$

the conditional distribution for α is given by

$$\alpha \mid \mathbf{w}_i, \mu, \sigma^2, \beta \sim N(M_\alpha, D_\alpha) I(|\alpha| < 1).$$

Sampling from this truncated normal distribution can be done using the method of inversion of the cumulative distribution function, as described in Robert (1995) and Gelfand et al. (1992). Simulate $u \sim U[0,1]$, then,

$$M_\alpha + D_\alpha^{1/2} \Phi^{-1} \left(u \left(\Phi\left(\frac{1 - M_\alpha}{D_\alpha^{1/2}} \right) - \Phi\left(\frac{-1 - M_\alpha}{D_\alpha^{1/2}} \right) \right) + \Phi\left(\frac{-1 - M_\alpha}{D_\alpha^{1/2}} \right) \right)$$

will be from the required truncated Normal distribution, where $\Phi()$ is the distribution function of the standard Normal distribution.

A.1.4 β

The hyperparameter β has prior distribution

$$\beta \sim \Gamma(g, h)$$

so that,

$$f(\beta) \propto \beta^{g-1} \exp[-h\beta].$$

The parameter β does not feature in the likelihood, but the prior distribution of σ^2 is dependent on β, so to obtain the conditional distribution of β, the prior distribution given above is combined

with the prior distribution of the parameter σ^2. The conditional density function is then given by

$$f(\beta \mid \sigma^2) \propto \beta^{g-1+\gamma} \exp\left[-\beta\left(h + \frac{1}{\sigma^2}\right)\right],$$

so the conditional distribution for β is of the form

$$\beta \mid \sigma^2 \sim \Gamma\left(\gamma + g, h + \frac{1}{\sigma^2}\right).$$

A.2 Hidden Markov model

In this section, the distributions of each of the nine parameters in $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \alpha, p_{01}, p_{10}, \beta_1, \beta_2)$, conditional on all of the remaining parameters, the general set of data $\mathbf{w}_i$ and the general set of hidden states $\mathbf{S}_i = (S_{i1}, \ldots, S_{in_{D_i}})$ are given. In addition, the method described in Chib (1996) for block sampling from the conditional distribution of the hidden states, given the data $\mathbf{w}_i$ and the parameter θ is discussed.

The Gibbs sampler is used to obtain draws from the posterior density function $f(\theta, \mathbf{S}_i \mid \mathbf{w}_i)$. It does this by sampling from the full conditionals $f(\theta_k \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-k})$ for $k \in \{1, \ldots, 9\}$ and $f(\mathbf{S}_i \mid \mathbf{w}_i, \theta)$. Using Bayes' theorem and the definition of a conditional probability density function, $f(\theta_k \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-k})$ can be written

$$\begin{aligned} f(\theta_k \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-k}) &\propto f(\theta, \mathbf{S}_i \mid \mathbf{w}_i) \\ &\propto f(\mathbf{w}_i \mid \theta, \mathbf{S}_i) f(\theta, \mathbf{S}_i) \\ &\propto f(\mathbf{w}_i \mid \theta, \mathbf{S}_i) f(\mathbf{S}_i \mid \theta) f(\theta) \\ &\propto f(\mathbf{w}_i \mid \theta, \mathbf{S}_i) f(\mathbf{S}_i \mid p_{01}, p_{10}) f(\theta). \end{aligned} \tag{A.3}$$

Denote the number of observations in state j by N_j, with $j \in \{1, 2, 3, 4\}$. The likelihood of the parameters and the hidden states, $f(\mathbf{w}_i \mid \theta, \mathbf{S}_i)$, or $L(\theta, \mathbf{S}_i)$ is given by

$$\begin{aligned} L(\theta, \mathbf{S}_i) \propto (2\pi\sigma_1^2)^{-\frac{N_1+N_3}{2}} (2\pi\sigma_2^2)^{-\frac{N_2+N_4}{2}} &\times \exp\left[-\frac{1}{2\sigma_1^2} I(S_{i1} = 1, 3)(w_{i1} - \mu_1)^2\right] \\ &\times \exp\left[-\frac{1}{2\sigma_2^2} I(S_{i1} = 2, 4)(w_{i1} - \mu_2)^2\right] \times \exp\left[-\sum_{t=2}^{n_{D_i}} \left(\frac{I(S_{it} = 1)}{2\sigma_1^2}(w_{it} - \mu_1 + \alpha\mu_1 - \alpha w_{i,t-1})^2\right.\right. \\ &+ \frac{I(S_{it} = 2)}{2\sigma_2^2}(w_{it} - \mu_2 + \alpha\mu_1 - \alpha w_{i,t-1})^2 + \frac{I(S_{it} = 3)}{2\sigma_1^2}(w_{it} - \mu_1 + \alpha\mu_2 - \alpha w_{i,t-1})^2 \\ &\left.\left. + \frac{I(S_{it} = 4)}{2\sigma_2^2}(w_{it} - \mu_2 + \alpha\mu_2 - \alpha w_{i,t-1})^2\right)\right], \end{aligned} \tag{A.4}$$

where $I(S_{it} = j)$ is the indicator function such that

$$I(S_{it} = j) = 1 \text{ if } S_{it} = j,$$
$$= 0 \text{ if } S_{it} \neq j.$$

The prior distributions of the parameters in θ are assumed independent, with the exception of σ_1^2 and β_1 and σ_2^2 and β_2. The prior density function $f(\theta)$ is therefore given by $f(\mu_1)f(\mu_2)f(\sigma_1^2 \mid \beta_1)f(\sigma_2^2 \mid \beta_2)f(\beta_1)f(\beta_2)f(\alpha)f(p_{01})f(p_{10})$. The forms of these individual prior density functions are given in Section 3.3.4. As discussed at the end of Section 3.3.6, the likelihood of the parameters is invariant under the relabelling of the states from $(1,2,3,4)$ to $(4,3,2,1)$, with an associated relabelling of the parameters associated with each of these states (so that e.g. μ_1 switches with μ_2). The prior distributions are also invariant under this same relabelling. As a result, the permutation sampler given in Frühwirth-Schnatter (2001), and detailed at the end of Section 3.3.6 should also be used as a final step after each full draw of θ from the Gibbs sampler.

Similar Gibbs sampling schemes for hidden Markov models can be seen in Albert and Chib (1993) and Kim and Nelson (1999), although both of these have slightly different forms to that given here because the form of the hidden Markov model is slightly different. A discussion of these differences is given in Section 3.3.3.

A.2.1 p_{01}, p_{10}

The transition probabilities have a beta prior distribution, truncated between $2/n_{D_i}$ and $(n_{D_i}-2)/n_{D_i}$, where n_{D_i} is the number of observations in the sample, so that

$$f(p_{01}) \propto (1-p_{01})^{b-1} p_{01}^{a-1} I\left(2/n_{D_i} \leq p_{01} \leq (n_{D_i}-2)/n_{D_i}\right)$$
$$f(p_{10}) \propto (1-p_{10})^{b-1} p_{10}^{a-1} I\left(2/n_{D_i} \leq p_{10} \leq (n_{D_i}-2)/n_{D_i}\right).$$

The likelihood does not contain p_{01} or p_{10}, but the function $f(\mathbf{S}_i \mid p_{01}, p_{10})$ in (A.3) does. By considering the transition matrix (3.12), this function can be seen to be equal to

$$f(\mathbf{S}_i \mid p_{01}, p_{10}) \propto (1-p_{01})^{n_{11}+n_{31}} p_{01}^{n_{12}+n_{32}} (1-p_{10})^{n_{24}+n_{44}} p_{10}^{n_{23}+n_{43}},$$

where n_{jk} with $j,k \in \{1,2,3,4\}$ indicates the number of transitions from state j to state k in the sample.

Combining the prior distributions and $f(\mathbf{S}_i \mid p_{01}, p_{10})$ gives

$$f(p_{01} \mid \mathbf{S}_i) \propto (1-p_{01})^{n_{11}+n_{31}+b-1} p_{01}^{n_{12}+n_{32}+a-1} I\left(2/n_{D_i} \leq p_{01} \leq (n_{D_i}-2)/n_{D_i}\right)$$
$$f(p_{10} \mid \mathbf{S}_i) \propto (1-p_{10})^{n_{24}+n_{44}+b-1} p_{10}^{n_{23}+n_{43}+a-1} I\left(2/n_{D_i} \leq p_{01} \leq (n_{D_i}-2)/n_{D_i}\right).$$

So, the conditional distributions for p_{01} and p_{10} are given by

$$p_{01} \mid \mathbf{S}_i \sim \text{Beta}(n_{12} + n_{32} + a, n_{11} + n_{31} + b) I\left(2/n_{D_i} \leq p_{01} \leq (n_{D_i} - 2)/n_{D_i}\right)$$
$$p_{10} \mid \mathbf{S}_i \sim \text{Beta}(n_{23} + n_{43} + a, n_{24} + n_{44} + b) I\left(2/n_{D_i} \leq p_{01} \leq (n_{D_i} - 2)/n_{D_i}\right).$$

The parameters p_{01} and p_{10} are drawn from a beta distribution, truncated to lie between $2/n_{D_i}$ and $(n_{D_i} - 2)/n_{D_i}$. To simulate from this distribution, a Metropolis Hastings step with a beta proposal distribution can be used. If the current value in the chain (at draw r) is given by $p_{01}^{(r)}$, a new value p'_{01} is proposed from the proposal distribution given by

$$p'_{01} \sim \text{Beta}(n_{12} + n_{32} + a, n_{11} + n_{31} + b).$$

If p'_{01} lies between $2/n_{D_i}$ and $(n_{D_i} - 2)/n_{D_i}$, the proposal is accepted, and hence $p_{01}^{(r+1)}$ is set to p'_{01}. Otherwise, set $p_{01}^{(r+1)} = p_{01}^{(r)}$. The same method is used to sample p_{10}.

A.2.2 μ_1, μ_2

The prior distributions of μ_1 and μ_2 are Normal so that

$$\mu_1 \sim N(\mu_0, V_\mu)$$
$$\mu_2 \sim N(\mu_0, V_\mu).$$

Combining the prior distribution of μ_1 with the likelihood (A.4), and discarding constant terms, the conditional density function of μ_1 is proportional to

$$\begin{aligned} f(\mu_1 \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-\mu_1}) \propto \exp\Bigg[& -\frac{1}{2V_\mu}(\mu_1 - \mu_0)^2 - \frac{I(S_{i1} = 1, 3)}{2\sigma_1^2}(w_{i1} - \mu_1)^2 \\ & - \sum_{t=2}^{n_{D_i}} \Bigg(\frac{I(S_{it} = 1)}{2\sigma_1^2}(w_{it} - \mu_1 + \alpha\mu_1 - \alpha w_{i,t-1})^2 \\ & + \frac{I(S_{it} = 2)}{2\sigma_2^2}(w_{it} - \mu_2 + \alpha\mu_1 - \alpha w_{i,t-1})^2 \\ & + \frac{I(S_{it} = 3)}{2\sigma_1^2}(w_{it} - \mu_1 + \alpha\mu_2 - \alpha w_{i,t-1})^2 \Bigg) \Bigg]. \end{aligned} \tag{A.5}$$

Letting,

$$A_{\mu_1} = \left[\frac{1}{V_\mu} + \frac{1}{\sigma_1^2} \sum_{t=2}^{n_{D_i}} \left[I(S_{it} = 1)(\alpha - 1)^2 + I(S_{it} = 3) \right] + \frac{1}{\sigma_2^2} \sum_{t=2}^{n_{D_i}} I(S_{it} = 2)\alpha^2 + \frac{1}{\sigma_1^2} I(S_{i1} = 1, 3) \right]^{-1}$$

and

$$M_{\mu_1} = A_{\mu_1}\left[\frac{\mu_0}{V_\mu} + \frac{1}{\sigma_1^2} I(S_{i1}=1,3)\, w_{i1} + \frac{1}{\sigma_1^2}\sum_{t=2}^{n_{D_i}}\big[I(S_{it}=1)(\alpha-1)(\alpha w_{i,t-1} - w_{it}) \\ + \; I(S_{it}=3)(w_{it} + \alpha\mu_2 - \alpha w_{i,t-1})\big] + \frac{1}{\sigma_2^2}\sum_{t=2}^{n_{D_i}} I(S_{it}=2)\alpha(\alpha w_{i,t-1} + \mu_2 - w_{it})\right],$$

equation (A.5) becomes

$$f(\mu_1 \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-\mu_1}) \propto \exp\left[-\frac{1}{2A_{\mu_1}}(\mu_1 - M_{\mu_1})^2\right]$$

and so the conditional distribution of μ_1 is given by

$$\mu_1 \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-\mu_1} \sim N(M_{\mu_1}, A_{\mu_1}).$$

Similarly, letting

$$A_{\mu_2} = \left[\frac{1}{V_\mu} + \frac{1}{\sigma_2^2}\sum_{t=2}^{n_{D_i}}\big[I(S_{it}=4)(\alpha-1)^2 + I(S_{it}=2)\big] + \frac{1}{\sigma_1^2}\sum_{t=2}^{n_{D_i}} I(S_{it}=3)\alpha^2 + \frac{1}{\sigma_2^2} I(S_{i1}=2,4)\right]^{-1}$$

$$M_{\mu_2} = A_{\mu_2}\left[\frac{\mu_0}{V_\mu} + \frac{1}{\sigma_2^2} I(S_{i1}=2,4)\, w_{i1} + \frac{1}{\sigma_2^2}\sum_{t=2}^{n_{D_i}}\big[I(S_{it}=4)(\alpha-1)(\alpha w_{i,t-1} - w_{it}) \\ + I(S_{it}=2)(w_{it} + \alpha\mu_1 - \alpha w_{i,t-1})\big] + \frac{1}{\sigma_1^2}\sum_{t=2}^{n_{D_i}} I(S_{it}=3)\alpha(\alpha w_{i,t-1} + \mu_1 - w_{it})\right],$$

the conditional distribution of μ_2 is given by

$$\mu_2 \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-\mu_2} \sim N(M_{\mu_2}, A_{\mu_2}).$$

A.2.3 σ_1^2, σ_2^2

The prior distributions of σ_1^2 and σ_2^2, given the hyperparameters β_1 and β_2, are inverse gamma with parameters

$$\sigma_1^2 \mid \beta_1 \sim IG(\gamma, \beta_1)$$
$$\sigma_2^2 \mid \beta_2 \sim IG(\gamma, \beta_2),$$

so that

$$f(\sigma_1^2 \mid \beta_1) \propto \sigma_1^{-(2\gamma+2)} \exp\left[-\frac{\beta_1}{\sigma_1^2}\right]$$
$$f(\sigma_2^2 \mid \beta_2) \propto \sigma_2^{-(2\gamma+2)} \exp\left[-\frac{\beta_2}{\sigma_2^2}\right].$$

Combining the prior distribution of σ_1^2 with the likelihood (A.4), the conditional density function of σ_1^2 is proportional to

$$\begin{aligned}
f(\sigma_1^2 \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-\sigma_1^2}) \propto \sigma_1^{-(2\gamma+2)} \exp\left[-\frac{\beta_1}{\sigma_1^2}\right] (2\pi\sigma_1^2)^{-\frac{N_1+N_3}{2}} \exp\left[\frac{-I(S_{i1}=1,3)}{2\sigma_1^2}(w_{i1}-\mu_1)^2\right. \\
\left. -\sum_{t=2}^{n_{D_i}}\left(\frac{I(S_{it}=1)}{2\sigma_1^2}(w_{it}-\mu_1+\alpha\mu_1-\alpha w_{i,t-1})^2\right.\right. \\
\left.\left. +\frac{I(S_{it}=3)}{2\sigma_1^2}(w_{it}-\mu_1+\alpha\mu_2-\alpha w_{i,t-1})^2\right)\right] \\
\propto \sigma_1^{-(2\gamma+2+N_1+N_3)} \exp\left[-\frac{1}{2\sigma_1^2}\left(2\beta_1 + I(S_{i1}=1,3)(w_{i1}-\mu_1)^2\right.\right. \\
\left.\left. +\sum_{t=2}^{n_{D_i}}\left(I(S_{it}=1)(w_{it}-\mu_1+\alpha\mu_1-\alpha w_{i,t-1})^2\right.\right.\right. \\
\left.\left.\left. +I(S_{it}=3)(w_{it}-\mu_1+\alpha\mu_2-\alpha w_{i,t-1})^2\right)\right)\right].
\end{aligned}$$

Letting

$$\begin{aligned}
D_{\sigma_1} = &\frac{1}{2}\left[2\beta_1 + I(S_{i1}=1,3)(w_{i1}-\mu_1)^2 + \sum_{t=2}^{n_{D_i}}(I(S_{it}=1)(w_{it}-\mu_1+\alpha\mu_1-\alpha w_{i,t-1})^2\right. \\
&\left. +I(S_{it}=3)(w_{it}-\mu_1+\alpha\mu_2-\alpha w_{i,t-1})^2)\right],
\end{aligned}$$

the conditional distribution of σ_1^2 is given by

$$\sigma_1^2 \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-\sigma_1^2} \sim IG\left(\frac{2\gamma+N_1+N_3}{2}, D_{\sigma_1}\right).$$

Similarly, the conditional distribution of σ_2^2 is given by

$$\sigma_2^2 \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-\sigma_2^2} \sim IG\left(\frac{2\gamma+N_2+N_4}{2}, D_{\sigma_2}\right),$$

where

$$\begin{aligned}
D_{\sigma_2} = &\frac{1}{2}\left[2\beta_2 + I(S_{i1}=2,4)(w_{i1}-\mu_2)^2 + \sum_{t=2}^{n_{D_i}}(I(S_{it}=4)(w_{it}-\mu_2+\alpha\mu_2-\alpha w_{i,t-1})^2\right. \\
&\left. +I(S_{it}=2)(w_{it}-\mu_2+\alpha\mu_1-\alpha w_{i,t-1})^2)\right].
\end{aligned}$$

A.2.4 α

The prior distribution of the autocorrelation parameter is Normal and restricted to lie between -1 and 1, so that

$$\alpha \sim N(\alpha_0, V_\alpha) I(|\alpha| < 1).$$

The likelihood (A.4) combines with this prior distribution to give a conditional density function proportional to

$$\begin{aligned} f(\alpha \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-\alpha}) \propto \exp\Bigg[-\frac{1}{2V_\alpha}(\alpha-\alpha_0)^2 - \sum_{t=2}^{n_{D_i}} \Bigg(& \frac{I(S_{it}=1)}{2\sigma_1^2}(w_{it}-\mu_1+\alpha\mu_1-\alpha w_{i,t-1})^2 \\ & + \frac{I(S_{it}=2)}{2\sigma_2^2}(w_{it}-\mu_2+\alpha\mu_1-\alpha w_{i,t-1})^2 \\ & + \frac{I(S_{it}=3)}{2\sigma_1^2}(w_{it}-\mu_1+\alpha\mu_2-\alpha w_{i,t-1})^2 \\ & + \frac{I(S_{it}=4)}{2\sigma_2^2}(w_{it}-\mu_2+\alpha\mu_2-\alpha w_{i,t-1})^2 \Bigg) \Bigg] I(|\alpha|<1). \end{aligned}$$

So letting

$$\begin{aligned} D_\alpha = & \left(\frac{1}{V_\alpha} + \sum_{t=2}^{n_{D_i}} \left(\frac{I(S_{it}=1)}{\sigma_1^2}(w_{i,t-1}-\mu_1)^2 + \frac{I(S_{it}=2)}{\sigma_2^2}(w_{i,t-1}-\mu_1)^2 \right.\right. \\ & \left.\left. + \frac{I(S_{it}=3)}{\sigma_1^2}(w_{i,t-1}-\mu_2)^2 + \frac{I(S_{it}=4)}{\sigma_2^2}(w_{i,t-1}-\mu_2)^2 \right) \right)^{-1} \\ M_\alpha = & D_\alpha \left(\frac{\alpha_0}{V_\alpha} + \sum_{t=2}^{n_{D_i}} \left(\frac{I(S_{it}=1)}{\sigma_1^2}(w_{i,t-1}-\mu_1)(w_{it}-\mu_1) \right.\right. \\ & + \frac{I(S_{it}=2)}{\sigma_2^2}(w_{i,t-1}-\mu_1)(w_{it}-\mu_2) + \frac{I(S_{it}=3)}{\sigma_1^2}(w_{i,t-1}-\mu_2)(w_{it}-\mu_1) \\ & \left.\left. + \frac{I(S_{it}=4)}{\sigma_2^2}(w_{i,t-1}-\mu_2)(w_{it}-\mu_2) \right) \right) \end{aligned}$$

the posterior density function is given by

$$f(\alpha \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-\alpha}) \propto \exp\left[-\frac{1}{2D_\alpha}(\alpha - M_\alpha)^2 \right].$$

Therefore, the conditional distribution of α is given by

$$\alpha \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-\alpha} \sim N(M_\alpha, D_\alpha) I(|\alpha|<1).$$

This is a truncated Normal distribution and it can be sampled from using the method described in Section A.1.3.

A.2.5 β_1, β_2

The prior distributions of β_1 and β_2 are given by

$$\beta_1, \beta_2 \sim \Gamma(g, h).$$

The conditional density functions $f(\beta_1 \mid \sigma_1^2)$ and $f(\beta_2 \mid \sigma_2^2)$ are derived as in Section A.1.4. They

are given by

$$\beta_1 \mid \sigma_1^2 \sim \Gamma\left(\gamma + g, h + \frac{1}{\sigma_1^2}\right)$$
$$\beta_2 \mid \sigma_2^2 \sim \Gamma\left(\gamma + g, h + \frac{1}{\sigma_2^2}\right).$$

A.2.6 Sampling the hidden states

In Albert and Chib (1993) and Chib (1996), Gibbs samplers were used to estimate the posterior distribution of the parameters and states of a hidden Markov model, conditional on the data. The hidden states are parameters in the model, and so the conditional probability density function $f(\mathbf{S}_i \mid \mathbf{w}_i, \theta)$ must be sampled from, as well as the conditional density functions $f(\theta_k \mid \mathbf{w}_i, \mathbf{S}_i, \theta_{-k})$, which were given in the previous sections for the particular model used here. In Albert and Chib (1993), the conditional density functions of each individual state, $f(S_{it} \mid S_{i,-t}, \theta, \mathbf{w}_i)$ were sampled from for all $t \in \{1, \ldots, n_{D_i}\}$. In Chib (1996) a procedure was introduced to sample from the joint density function $f(\mathbf{S}_i \mid \theta, \mathbf{w}_i)$ directly. Block sampling all of the states in one step in this way should improve the mixing of the chain, so this is the method discussed here.

The method presented in Chib (1996) for obtaining draws from $f(\mathbf{S}_i \mid \theta, \mathbf{w}_i)$ will be summarised. Following the notation used in that paper, let $\mathbf{S}_{it} = (S_{i1}, \ldots, S_{it})$ and $\mathbf{S}_i^{t+1} = (S_{i,t+1}, \ldots, S_{i,n_{D_i}})$, with similar notation used for $\mathbf{w}_{it}$ and $\mathbf{w}_i^{t+1}$. The t-th single state will be denoted S_{it}.

The required joint density function of the states is rewritten as

$$\begin{aligned} f(\mathbf{S}_i \mid \theta, \mathbf{w}_i) &= f(S_{i1}, \ldots, S_{in_{D_i}} \mid \theta, \mathbf{w}_{in_{D_i}}) \\ &= f(S_{i1} \mid \theta, \mathbf{w}_{in_{D_i}}, \mathbf{S}_i^2) f(\mathbf{S}_i^2 \mid \theta, \mathbf{w}_{in_{D_i}}) \\ &= f(S_{i1} \mid \theta, \mathbf{w}_{in_{D_i}}, \mathbf{S}_i^2) f(S_{i2} \mid \theta, \mathbf{w}_{in_{D_i}}, \mathbf{S}_i^3) \ldots f(S_{in_{D_i}} \mid \theta, \mathbf{w}_{in_{D_i}}). \end{aligned}$$

So, it is possible to sample from $f(\mathbf{S}_i \mid \theta, \mathbf{w}_i)$ by sampling from each of $f(S_{it} \mid \theta, \mathbf{w}_{in_{D_i}}, \mathbf{S}_i^{t+1})$ for all $t \in \{1, \ldots, n_{D_i}\}$, starting with $t = n_{D_i}$.

Using Bayes' theorem, each of the terms $f(S_{it} \mid \theta, \mathbf{w}_{in_{D_i}}, \mathbf{S}_i^{t+1})$ can be written as

$$\begin{aligned} f(S_{it} \mid \theta, \mathbf{w}_{in_{D_i}}, \mathbf{S}_i^{t+1}) &= f(S_{it} \mid \theta, \mathbf{w}_{it}, \mathbf{w}_i^{t+1}, \mathbf{S}_i^{t+1}) \\ &\propto f(\mathbf{w}_i^{t+1}, \mathbf{S}_i^{t+1} \mid \theta, \mathbf{w}_{it}, S_{it}) f(S_{it} \mid \theta, \mathbf{w}_{it}) \\ &\propto f(\mathbf{w}_i^{t+1}, \mathbf{S}_i^{t+2}, S_{i,t+1} \mid \theta, \mathbf{w}_{it}, S_{it}) f(S_{it} \mid \theta, \mathbf{w}_{it}) \\ &\propto f(\mathbf{w}_i^{t+1}, \mathbf{S}_i^{t+2} \mid \theta, \mathbf{w}_{i,t}, S_{it}, S_{i,t+1}) f(S_{i,t+1} \mid \theta, \mathbf{w}_{it}, S_{it}) f(S_{it} \mid \theta, \mathbf{w}_{it}). \end{aligned} \tag{A.6}$$

Equation (A.6) can be simplified using the conditional independence structure of the Bayesian network. The first term $f(\mathbf{w}_i^{t+1}, \mathbf{S}_i^{t+2} \mid \theta, \mathbf{w}_{it}, S_{it}, S_{i,t+1})$ can be simplified to $f(\mathbf{w}_i^{t+1}, \mathbf{S}_i^{t+2} \mid \theta, \mathbf{w}_{it}, S_{i,t+1})$, as $\mathbf{w}_i^{t+1}$ and $\mathbf{S}_i^{t+2}$ are not dependent on S_{it}, given $S_{i,t+1}$ and $\mathbf{w}_{it}$. Hence, the first term does not contain S_{it} and so it is constant. The term $f(S_{i,t+1} \mid \theta, \mathbf{w}_{it}, S_{it})$ can be simplified to $f(S_{i,t+1} \mid \theta, S_{it})$ because the function is not conditioned on the values of $w_{i,t+1}$, and so $S_{i,t+1}$ is conditionally independent of $\mathbf{w}_{it}$ given S_{it}. This is because the nodes of $S_{i,t+1}$, $\mathbf{w}_{it}$ and $w_{i,t+1}$ in the Bayesian network in figure 3.2 form a converging connection, and so the nodes $S_{i,t+1}$ and $\mathbf{w}_{it}$ are d-separated, and hence independent, when $w_{i,t+1}$, and all of $\mathbf{w}_i^{t+2}$, are not known. For more information on this, see p. 41 of Taroni, Aitken et al. (2006).

Following these simplifications, (A.6) becomes

$$f(S_{it} \mid \theta, \mathbf{w}_{in_{D_i}}, \mathbf{S}_i^{t+1}) \propto f(S_{i,t+1} \mid \theta, S_{it}) f(S_{it} \mid \theta, \mathbf{w}_{it}). \tag{A.7}$$

The term $f(S_{i,t+1} \mid \theta, S_{it})$ in this expression is given by the transition matrix of the hidden states, (3.12). Therefore, it remains to find an expression for $f(S_{it} \mid \theta, \mathbf{w}_{it})$ so that draws can be obtained from $f(S_{it} \mid \theta, \mathbf{w}_{in_{D_i}}, \mathbf{S}_i^{t+1})$, as required. Using Bayes' theorem, the second term in (A.7) can be written

$$\begin{aligned} f(S_{it} \mid \theta, \mathbf{w}_{it}) &= f(S_{it} \mid \theta, \mathbf{w}_{i,t-1}, w_{it}) \\ &\propto f(w_{it} \mid S_{it}, \theta, \mathbf{w}_{i,t-1}) f(S_{it} \mid \theta, \mathbf{w}_{i,t-1}). \end{aligned} \tag{A.8}$$

The first term in this expression can be evaluated using (3.13). It is the Normal probability density function, with mean $\mu_{S_{it}}^{(1)} + \alpha(w_{i,t-1} - \mu_{S_{it}}^{(2)})$ and variance $\sigma_{S_{it}}^2$. Using the law of total probability and Bayes' theorem, the second term can be written

$$\begin{aligned} f(S_{it} \mid \theta, \mathbf{w}_{i,t-1}) &= \sum_{S_{i,t-1} \in \{1,2,3,4\}} f(S_{i,t-1}, S_{it} \mid \theta, \mathbf{w}_{i,t-1}) \\ &= \sum_{S_{i,t-1} \in \{1,2,3,4\}} f(S_{i,t-1} \mid \theta, \mathbf{w}_{i,t-1}) f(S_{it} \mid S_{i,t-1}, \theta, \mathbf{w}_{i,t-1}) \\ &= \sum_{S_{i,t-1} \in \{1,2,3,4\}} f(S_{i,t-1} \mid \theta, \mathbf{w}_{i,t-1}) f(S_{it} \mid S_{i,t-1}, \theta). \end{aligned} \tag{A.9}$$

The second term in (A.9) is given by the transition matrix of the hidden states, (3.12). As stated in Chib (1996), the first term, for $t = 2$, can be obtained from the initial distribution of the hidden states. This initial distribution gives the four values of $f(S_{i1} \mid \theta, \mathbf{w}_{i0})$ for each possible value of S_{i1}, which can in turn be used to obtain $f(S_{i1} \mid \theta, \mathbf{w}_{i1})$, using (A.8).

Beginning with $f(S_{i1} \mid \theta, \mathbf{w}_{i1})$, each of $f(S_{it} \mid \theta, \mathbf{w}_{it})$ for $t \in \{2, \ldots, n_{D_i}\}$ can be calculated and stored by using first the expression (A.9) and then substituting the result into (A.8). Each of these functions is a discrete probability mass function giving the probabilities of being in each of the four states,

conditional on the parameters and the subset of data $\mathbf{w}_{it}$. By sampling from the distribution given by the probability mass function $f(S_{in_{D_i}} \mid \theta, \mathbf{w}_{in_{D_i}})$, a simulated state for the final observation is obtained. From this simulated final state, (A.7) (normalised) can be used to simulate the $(n_{D_i}-1)$-th state, using the stored probabilities $f(S_{i,n_{D_i}-1} \mid \theta, \mathbf{w}_{i,n_{D_i}-1})$ multiplied by the probabilities (from the transition matrix, (3.12)) of moving from each of the four possible states to the simulated n_{D_i}-th state. Equation (A.7) will then give the probabilities of the $(n_{D_i}-1)$-th state being each of $\{1,2,3,4\}$, conditional on the parameters, the data and the n_{D_i}-th state, so this probability mass function can be sampled from to obtain a simulated state for the $(n_{D_i}-1)$-th observation. Continuing in this way, a sample of the hidden states, dependent on the parameters and the data is obtained.

Appendix B

Calculation of marginal likelihoods for model selection

The i-th sample from a general set $\mathbf{w}$ is denoted $\mathbf{w}_i$. There are m_D such samples, and each sample has n_{D_i} observations. Let $\mathbf{M} = (M_1, \ldots, M_{m_D})$ denote the model choice for each of those m_D samples. Let $M_i = M_H$ if the model choice for the i-th sample is the hidden Markov model, and let $M_i = M_A$ if the model choice for the i-th sample is the autoregressive model of order one. In this section, methods for calculating the marginal likelihoods $f(\mathbf{w}_i \mid M_i = M_A)$ and $f(\mathbf{w}_i \mid M_i = M_H)$ are described, so that the Bayes factor

$$\frac{f(\mathbf{w}_i \mid M_i = M_A)}{f(\mathbf{w}_i \mid M_i = M_H)}$$

can be calculated. These methods are used in Section 4.4 of Chapter 4 to decide which of the two models (autoregressive or hidden Markov) to use for each sample. The two methods discussed are a straightforward Monte Carlo integration approach and a method developed in Chib and Jeliazkov (2001), which uses posterior draws from a Metropolis-Hastings sampler.

B.1 Monte Carlo integration

Monte Carlo integration can be used to calculate an estimate of the marginal likelihood $f(\mathbf{w}_i \mid M_i)$. This method is the same as that described in Chapter 4 for the estimation of the likelihood ratio. Monte Carlo integration is more inefficient than the method presented in the next section (Newton and Raftery (1994)), but it can be used when no draws from the posterior density function $f(\theta \mid M_i, \mathbf{w}_i)$ have been obtained.

Let the r-th draw from the prior distribution of θ be denoted by $\theta^{(r)}$ where $r \in \{1, \ldots, N\}$. The prior distributions for the autoregressive model are given in Section 3.1.1, and those for the hidden Markov model are given in Section 3.3.4. The function $f(\mathbf{w}_i \mid M_i)$ is given by

$$f(\mathbf{w}_i \mid M_i) = \int_\theta f(\mathbf{w}_i \mid M_i, \theta) f(\theta \mid M_i) \mathrm{d}\theta.$$

This can be written as

$$f(\mathbf{w}_i \mid M_i) = \mathbb{E}_{f(\theta|M_i)} \left[f(\mathbf{w}_i \mid M_i, \theta) \right]$$

which can be estimated by

$$f(\mathbf{w}_i \mid M_i) \approx (1/N) \sum_{r=1}^{N} f(\mathbf{w}_i \mid \theta^{(r)}, M_i).$$

The function $f(\mathbf{w}_i \mid \theta^{(r)}, M_i)$ is the likelihood of θ and M_i. For the autoregressive model ($M_i = M_A$), it is a product of Normal probability density functions (as given in (3.4)), and for the hidden Markov model ($M_i = M_H$) it can be calculated using the method given in Section 3.3.5.

The draws $\theta^{(r)}$, used to estimate the marginal likelihood, are obtained from the prior distribution of θ, so if this prior distribution is not concentrated around the areas of high likelihood, then a large number of these draws will have small likelihoods. As a result, a large number of draws may be required to obtain a good estimate of the marginal likelihood if the prior distribution of θ is not concentrated around the areas of high likelihood. For more information on this, see McCulloch and Rossi (1992).

B.2 Chib and Jeliazkov's method

Chib and Jeliazkov's method (Chib and Jeliazkov (2001)) uses draws obtained from the Metropolis-Hastings sampler discussed in Sections 3.1.2 and 3.3.6 to estimate the marginal likelihood $f(\mathbf{w}_i \mid M_i)$. As discussed in Frühwirth-Schnatter (2004), Chib and Jeliazkov's method suffers from problems if the Metropolis-Hastings sampler does not visit all of the modes of the posterior distribution in the correct relative proportions. The likelihood of the parameters of a hidden Markov model is often invariant under certain permutations of the states, so if the prior distribution of these parameters is also invariant under these permutations of the states, a fully mixing sample from the posterior distribution of the parameters should include draws from each of the possible permutations. As described in Section 3.3.6, the likelihood and prior distribution of the parameters of the hidden Markov model are both invariant under the transformation from $\theta^D_{H_i} = \left(\mu^D_{1_i}, \mu^D_{2_i}, (\sigma^D_{1_i})^2, (\sigma^D_{2_i})^2, \alpha^D_{1_i}, p^D_{01_i}, p^D_{10_i}, \beta^D_{1_i}, \beta^D_{2_i}\right)$ to $\theta^D_{H_i} = \left(\mu^D_{2_i}, \mu^D_{1_i}, (\sigma^D_{2_i})^2, (\sigma^D_{1_i})^2, \alpha^D_{1_i}, p^D_{10_i}, p^D_{01_i}, \beta^D_{2_i}, \beta^D_{1_i}\right)$, which corresponds to the switching of the state labels from $(1,2,3,4)$ to $(4,3,2,1)$. As such, the marginal posterior density functions of each of the parameters, with the exception of α_{1_i}, should be bimodal, if there is some separation between the parameters of different states. The final step in the Metropolis-Hastings algorithm in Section 3.3.6 is a permutation sampler, which randomly switches the labelling of the states after every draw. As a result, it is ensured that both possible permutations of the states are visited by the Metropolis-Hastings

sampler in a balanced fashion. Hence, the problems discussed in Frühwirth-Schnatter (2004) are not an issue for the sampler for the hidden Markov model given in Section 3.3.6, because the two posterior modes are both explored. Therefore, Chib and Jeliazkov's method can be used to estimate the marginal likelihood.

Using Bayes' theorem the marginal likelihood can be written

$$f(\mathbf{w}_i \mid M_i) = \frac{f(\mathbf{w}_i \mid \theta, M_i) f(\theta \mid M_i)}{f(\theta \mid M_i, \mathbf{w}_i)}, \tag{B.1}$$

where, if $M_i = M_A$ then $\theta = \theta^D_{A_i} = (\mu^D_i, (\sigma^D_i)^2, \alpha^D_i, \beta^D_i)$ and if $M_i = M_H$ then $\theta = \theta^D_{H_i} = (\mu^D_{1_i}, \mu^D_{2_i}, (\sigma^D_{1_i})^2, (\sigma^D_{2_i})^2, \alpha^D_{1_i}, p^D_{01_i}, p^D_{10_i}, \beta^D_{1_i}, \beta^D_{2_i})$. The left hand side of (B.1) is not a function of θ, so the right hand side can be evaluated for a specific value of θ. Taking logarithms, rearranging, and evaluating at the point $\theta = \theta^*$, (B.1) becomes

$$\log(f(\mathbf{w}_i \mid M_i)) = \log(f(\mathbf{w}_i \mid \theta^*, M_i)) + \log(f(\theta^* \mid M_i)) - \log(f(\theta^* \mid M_i, \mathbf{w}_i)). \tag{B.2}$$

In Sections 3.1.2 (autoregressive model) and 3.3.6 (hidden Markov model), transformations were made to some of the parameters to improve the mixing of the Metropolis-Hastings sampler. Let the function $h(\theta)$ denote this transformation. For the autoregressive model, $h(\theta) = (\mu, \log(\sigma^2), \alpha, \log(\beta))$, and for the hidden Markov model, $h(\theta) = (\mu_1, \mu_2, \log(\sigma_1^2), \log(\sigma_2^2), \alpha_1, \log(p_{01}/(1-p_{01})), \log(p_{10}/(1-p_{10})), \log(\beta_1), \log(\beta_2))$. The posterior density function of the transformed parameters $h(\theta)$ is given by

$$f(h(\theta) \mid M_i, \mathbf{w}_i) = f(\theta \mid M_i, \mathbf{w}_i) J(\theta) = \frac{f(\mathbf{w}_i \mid \theta, M_i) f(\theta \mid M_i) J(\theta)}{f(\mathbf{w}_i \mid M_i)},$$

where the function $J(\theta)$ is the Jacobian of the inverse of the transformation h. Therefore, analogously to the derivation of (B.2), the marginal log-likelihood in terms of the posterior density function of the transformed parameters $h(\theta)$ is given by

$$\log(f(\mathbf{w}_i \mid M_i)) = \log(f(\mathbf{w}_i \mid \theta^*, M_i)) + \log(f(\theta^* \mid M_i)) + \log(J(\theta^*)) - \log(f(h(\theta^*) \mid M_i, \mathbf{w}_i)). \tag{B.3}$$

The first term on the right hand side of (B.3) is equal to the log-likelihood of θ, evaluated at θ^*. For the autoregressive model ($M_i = M_A$), it is the logarithm of a product of Normal probability density functions (as given in (3.4)), and for the hidden Markov model ($M_i = M_H$) it can be calculated using the method given in Section 3.3.5. The second term on the right hand side is given by the prior density function, evaluated at the point θ^*. The prior distribution functions for the autoregressive model parameters are given in Section 3.1.1. The joint prior density function is given by the product of the density functions associated with the distributions given in Section 3.1.1 and is explicitly given in (3.5). The prior distribution functions for the hidden Markov model parameters are given in Section 3.3.4, and the joint density function associated with these prior distributions is given in (3.16). For the

autoregressive model, the Jacobian $J(\theta)$ is given by

$$J(\theta) = \sigma^2 \beta.$$

For the hidden Markov model, the Jacobian $J(\theta)$ is given by

$$J(\theta) = \sigma_1^2 \sigma_2^2 \beta_1 \beta_2 (1 - p_{01}) p_{01} (1 - p_{10}) p_{10}.$$

To obtain the marginal likelihood $f(\mathbf{w}_i \mid M_i)$ in (B.3), it is therefore only required to evaluate the posterior density function $f(h(\theta) \mid M_i, \mathbf{w}_i)$ at the point $\theta = \theta^*$.

Chib and Jeliazkov (2001) derive an estimator of $f(h(\theta^*) \mid M_i, \mathbf{w}_i)$ which uses the draws from the Metropolis-Hastings sampler $\theta^{(r)}$ for $r \in \{1, \ldots, N\}$ (if the model in question is the autoregressive model, then the $\theta^{(r)}$ arise from the sampler detailed in Section 3.1.2 and if the model being considered is the hidden Markov model then they arise from the sampler in Section 3.3.6). The estimator is given by

$$\hat{f}(h(\theta^*) \mid \mathbf{w}_i, M_i) = \frac{\frac{1}{N} \sum_{r=1}^{N} A(\theta^{(r)}, \theta^* \mid \mathbf{w}_i, M_i) q(h(\theta^{(r)}), h(\theta^*) \mid M_i)}{\frac{1}{G} \sum_{g=1}^{G} A(\theta^*, \theta^{(g)} \mid \mathbf{w}_i, M_i)}. \tag{B.4}$$

The function $A(\theta, \theta' \mid \mathbf{w}_i, M_i)$ is the acceptance probability for the Metropolis-Hastings sampler. It is given in (3.6) for the autoregressive model and in (3.17) for the hidden Markov model. Note that this function is given in terms of the untransformed parameter θ. The function $q(h(\theta), h(\theta') \mid M_i)$ is the proposal distribution for the Metropolis-Hastings sampler. The proposal distributions for the two models are given in Sections 3.1.2 and 3.3.6, and are multivariate Normal distributions. The function $q(h(\theta), h(\theta') \mid M_i)$ is given, in both cases, by

$$q(h(\theta), h(\theta') \mid M_i) = (2\pi)^{-d/2} |V|^{-1/2} \exp\left[-\frac{1}{2} (h(\theta) - h(\theta'))^T V^{-1} (h(\theta) - h(\theta')) \right]$$

where d is 4 for the autoregressive model and 9 for the hidden Markov model. The $d \times d$ covariance matrix V is diagonal, with entries $V_1, \ldots, V_k$, where the values of V_k are the proposal variances of the Metropolis-Hastings samplers, as defined in Sections 3.1.2 and 3.3.6. The draws $\theta^{(g)}$ for $g \in \{1, \ldots, G\}$ in (B.4) are inversely transformed draws from the proposal distribution $q(h(\theta^*), h(\theta) \mid M_i)$, given the fixed value θ^*.

The estimator in (B.4) can be inserted in place of $f(h(\theta^*) \mid M_i, \mathbf{w}_i)$ in (B.3) to give an estimate of the required marginal likelihood $f(\mathbf{w}_i \mid M_i)$.

Appendix C

JAGS model code for autoregressive model with random effects

The JAGS code used with package rjags in R (Plummer (2013)) to obtain posterior draws for the autoregressive model with random effects in section 6.2.2 is given by:

```
model{
for(i in 1:mC){
mu[i]~dnorm(mu.mu, tau.mu)
alpha[i]~dnorm(mu.alpha,tau.alpha)T(-1,1)
tau[i]~dgamma(gamma.V,beta.V)
w[i,1]~dnorm(mu[i], tau[i])
for(t in 2:t.max[i]){
w[i,t]~dnorm(mu[i]+alpha[i]*(w[i,t-1]-mu[i]), tau[i])
}
}

mu.mu~dnorm(6, 0.05)
tau.mu<-pow(sigma.mu, -2)
sigma.mu~dunif(0,5)
gamma.V~dgamma(0.1,0.1)
beta.V~dgamma(0.1,0.1)
mu.alpha~dnorm(0,0.25)
tau.alpha<-pow(sigma.alpha, -2)
sigma.alpha~dunif(0,5)
}
```

Appendix D

Summary of samples and exhibits included in datasets B and C

Exhibit number	Case number	No. of notes	Exhibit number	Case number	No. of notes	Exhibit number	Case number	No. of notes
1	1	46	26	13	39	51	23	104
2	1	79	27	13	139	52	24	57
3	2	32	28	14	193	53	24	145
4	3	132	29	14	94	54	25	135
5	3	96	30	14	24	55	25	48
6	4	22	31	14	35	56	26	37
7	5	276	32	15	125	57	26	21
8	5	71	33	15	104	58	27	140
9	5	189	34	16	98	59	28	164
10	5	38	35	16	211	60	28	40
11	6	62	36	16	83	61	28	20
12	6	101	37	16	35	62	3	303
13	6	125	38	17	21	63	29	352
14	6	73	39	17	100	64	29	20
15	6	63	40	18	303	65	29	39
16	6	50	41	19	92	66	29	168
17	7	58	42	19	47	67	9	1099
18	8	47	43	19	91	68	9	1065
19	9	1030	44	20	41	69	9	1023
20	10	606	45	21	35	70	9	26
21	10	70	46	21	129			
22	11	237	47	21	75			
23	11	34	48	21	107			
24	11	148	49	21	51			
25	12	151	50	22	38			

Table D.1: Dataset *C*- exhibit numbers as referred to in the text, with case number (a case consists of multiple exhibits from the same criminal case) and the estimated number of banknotes in each exhibit. The estimated number of banknotes is determined by the number of peaks detected by the peak detection algorithm.

Sample number	No. of notes	Sample number	No. of notes	Sample number	No. of notes	Sample number	No. of notes
1	150	51	26	101	257	151	150
2	77	52	100	102	149	152	163
3	75	53	187	103	149	153	133
4	73	54	113	104	149	154	152
5	77	55	151	105	124	155	211
6	77	56	21	106	149	156	149
7	100	57	21	107	150	157	147
8	154	58	21	108	37	158	140
9	99	59	99	109	46	159	71
10	94	60	108	110	53	160	201
11	97	61	128	111	49	161	147
12	119	62	152	112	149	162	156
13	221	63	153	113	142	163	144
14	124	64	153	114	197	164	150
15	113	65	152	115	146	165	137
16	120	66	149	116	150	166	75
17	142	67	146	117	101	167	152
18	139	68	151	118	73	168	155
19	118	69	158	119	195	169	100
20	128	70	115	120	141	170	141
21	125	71	98	121	203	171	150
22	75	72	166	122	124	172	155
23	124	73	186	123	194	173	145
24	120	74	126	124	70	174	155
25	139	75	117	125	159	175	146
26	149	76	109	126	150	176	156
27	120	77	94	127	145	177	185
28	122	78	109	128	49	178	79
29	123	79	106	129	94	179	154
30	123	80	109	130	43	180	149
31	145	81	134	131	118	181	117
32	144	82	138	132	128	182	120
33	122	83	137	133	158	183	155
34	95	84	145	134	149	184	34
35	156	85	132	135	134	185	152
36	117	86	168	136	146	186	146
37	122	87	109	137	136	187	131
38	77	88	109	138	134	188	154
39	114	89	108	139	113	189	154
40	101	90	107	140	100	190	152
41	96	91	109	141	151	191	154
42	148	92	109	142	158	192	112
43	118	93	109	143	147	193	154
44	144	94	108	144	150		
45	86	95	125	145	109		
46	21	96	166	146	153		
47	21	97	151	147	144		
48	115	98	148	148	162		
49	21	99	152	149	134		
50	27	100	144	150	102		

Table D.2: Dataset *B*- sample numbers as referred to in the text, with the estimated number of banknotes in the sample. The estimated number of banknotes is determined by the number of peaks detected by the peak detection algorithm.